

Tvorba LLM (Large Language Models) v jazyku Python

Kód kurzu: PYTHON_LLM

Kurz je určený pre programátorov, ktorí chcú pochopiť, ako veľké jazykové modely (LLM) fungujú „zvnútra“, a prakticky si prejsť cestu od jednoduchých štatistických jazykových modelov cez RNN/LSTM až po transformer a mini GPT. V druhej časti kurzu sa účastníci naučia pracovať s hotovými modelmi (Hugging Face), vykonať fine-tuning (vrátane LoRA) a postaviť praktickú aplikáciu nad vlastnými dokumentmi pomocou RAG (retrieval-augmented generation). Súčasťou sú aj produkčné aspekty: latencia, optimalizácia, kvantizácia a nasadenie ako API vrátane Dockeru.

Pro koho je kurz určený:

- Programátorov so základnou znalosťou Pythonu, ktorí chcú porozumieť LLM a vedieť ich prakticky využiť.
- Vývojárov, ktorí chcú implementovať a trénovať malé modely a následne pracovať s existujúcimi LLM (fine-tuning, RAG, nasadenie).
- Data/ML nadšencov, ktorí chcú získať ucelený praktický prehľad od „nuly“ po aplikáciu

Požadované vstupné znalosti

- Základná znalosť jazyka Python na úrovni kurzu PYTHON_INTRO
- Základná znalosť NumPy na úrovni kurzu PYTHON_DATAN
- Základy lineárnej algebry sú výhodou (nie sú podmienkou)

Metódy výučby

- Odborný výklad s praktickými ukážkami, cvičenia na počítačoch.
- Práca v Jupyter Notebook a skriptoch, priebežné miniprojekty.
- Dôraz na pochopenie princípov a reprodukovateľnosť.

Študijné materiály

- Prezentácia preberanej látky v tlačenej alebo online forme.
- Notebooky a referenčné implementácie k jednotlivým dňom.
- Ukážkové datasety a šablóny pre tréning, vyhodnotenie a nasadenie.

Osnova

Základy neurónových sietí a NLP

- Čo je jazykový model
- Pravdepodobnosť a predikcia ďalšieho slova / tokenu
- Tokenizácia
- Embeddingy
- Neurónová sieť (perceptrón, vrstvy, aktivácie)
- Backpropagation (intuícia)
- Python + NumPy
- Implementácia unigram modelu
- Implementácia bigram modelu
- Tréning malej siete v PyTorch
- Výstup dňa: malý jazykový model predikujúci ďalšie slovo na malých vzorkách textu

Rekurentné siete a prvé generovanie textu

- RNN
- LSTM
- Problém miznúceho gradientu
- Teacher forcing
- Sampling (temperature, top-k)
- Implementácia LSTM modelu v PyTorch
- Tréning na malom datasete (napr. Shakespeare)

GOPAS Praha

Kodaňská 1441/46
101 00 Praha 10
Tel.: +420 234 064 900-3
info@gopas.cz

GOPAS Brno

Nové sady 996/25
602 00 Brno
Tel.: +420 542 422 111
info@gopas.cz

GOPAS Bratislava

Dr. Vladimíra Clementisa 10
Bratislava, 821 02
Tel.: +421 248 282 701-2
info@gopas.sk



Copyright © 2020 GOPAS, a.s.,
All rights reserved

Tvorba LLM (Large Language Models) v jazyku Python

- Generovanie textu
- Výstup dňa: model generujúci krátke texty

Transformer architektúra

- Attention mechanizmus
- Self-attention
- Multi-head attention
- Positional encoding
- Encoder vs Decoder
- Prečo je transformer škálovateľný
- Implementácia mini-transformera
- Vytvorenie malého GPT-like modelu
- Tréning na malom datasete
- Výstup dňa: funkčný mini GPT model

Tréning, fine-tuning a práca s hotovými modelmi

- Pretraining vs Fine-tuning
- Transfer learning
- LoRA a parameter-efficient fine-tuning (PEFT)
- Tokenizéry (BPE)
- Použitie knižnice Hugging Face Transformers
- Fine-tuning malého modelu
- Práca s modelmi (napr. LLaMA / kompatibilné open-source modely podľa dostupnosti)
- Vytvorenie vlastného chatbot skriptu
- Výstup dňa: fine-tuned model na vlastných dátach

RAG, nasadenie a produkčné aspekty

- Embeddingy pre vyhľadávanie
- Vektorové databázy (FAISS)
- RAG architektúra
- Latencia a optimalizácia
- Kvantizácia modelov
- Nasadenie (API, Docker)
- Generovanie embeddingov
- Uloženie do FAISS
- Implementácia RAG pipeline
- Vytvorenie jednoduchého API (FastAPI)
- Finálny projekt: interný chatbot nad vlastnými dokumentmi

Používané technológie

- Python 3.11+
- PyTorch
- Hugging Face Transformers
- FAISS
- FastAPI
- Jupyter Notebook

Čo si účastník odnesie

- Rozumie, ako LLM fungujú vnútorne a aké sú ich stavebné bloky
- Vie postaviť jednoduchý transformer a mini GPT na malom datasete
- Vie fine-tunovať model vrátane LoRA/PEFT
- Vie postaviť RAG aplikáciu nad vlastnými dokumentmi

GOPAS Praha

Kodaňská 1441/46
101 00 Praha 10
Tel.: +420 234 064 900-3
info@gopas.cz

GOPAS Brno

Nové sady 996/25
602 00 Brno
Tel.: +420 542 422 111
info@gopas.cz

GOPAS Bratislava

Dr. Vladimíra Clementisa 10
Bratislava, 821 02
Tel.: +421 248 282 701-2
info@gopas.sk



Copyright © 2020 GOPAS, a.s.,
All rights reserved

Tvorba LLM (Large Language Models) v jazyku Python

- Vie nasadiť model ako API a rozumie základným produkčným aspektom (latencia, optimalizácia, kvantizácia)

GOPAS Praha
Kodaňská 1441/46
101 00 Praha 10
Tel.: +420 234 064 900-3
info@gopas.cz

GOPAS Brno
Nové sady 996/25
602 00 Brno
Tel.: +420 542 422 111
info@gopas.cz

GOPAS Bratislava
Dr. Vladimíra Clementisa 10
Bratislava, 821 02
Tel.: +421 248 282 701-2
info@gopas.sk



Copyright © 2020 GOPAS, a.s.,
All rights reserved