Conversational Image Understanding with vision models

Arsenii Milenchuk, Dariia Shevchuk, Nazarii Tymochko, Oleksandr Petsa

KKUI / FEI / TUKE

2025

Definícia úlohy

- Prehľad literatúry - ✔
- Pochopenie architektúry - ✔
- Rozhranie chatu na nadväzovanie konverzácií s large language modelom - ✔ *už bol ako pet-projekt v minulosti*
- Vyhodnotenie rôznych veľkých multimodálnych modelov - ✔
- Pridávanie nástrojov pre použitie multimodalnym modelom - ✘
- Dokumentácia - ✘

Dataset

Kaggle

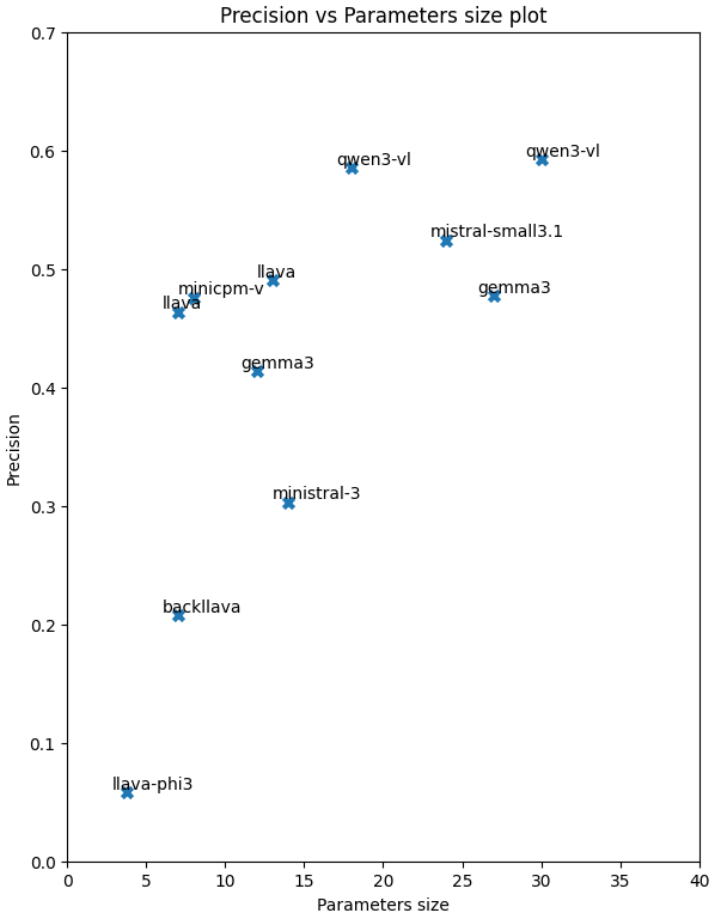
- Visual Question Answering-Computer Vision & NLP

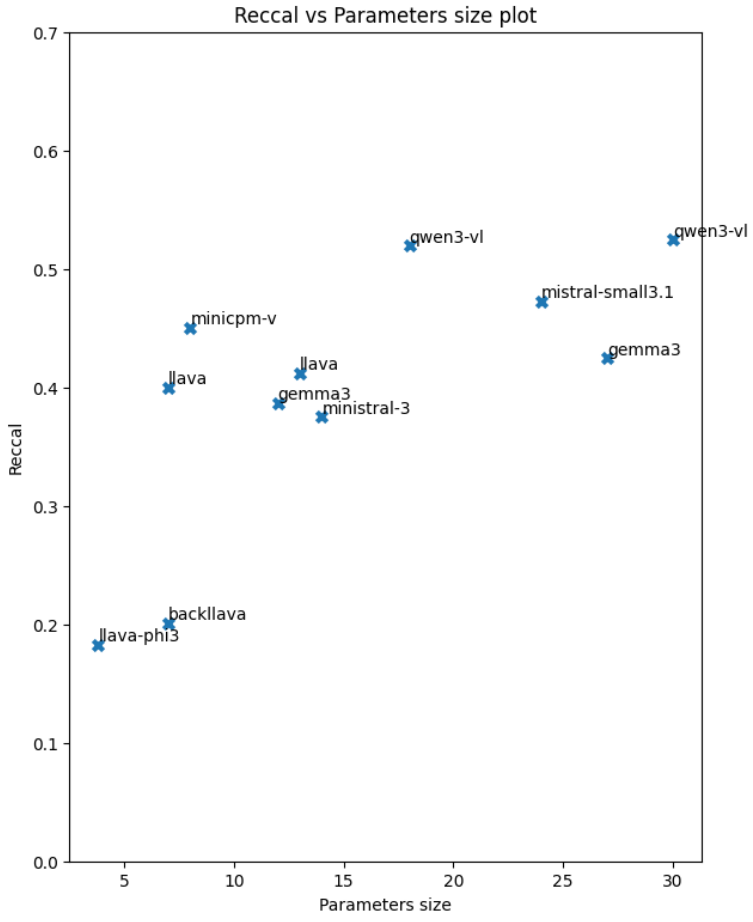
Hodnotenie

- BertScore

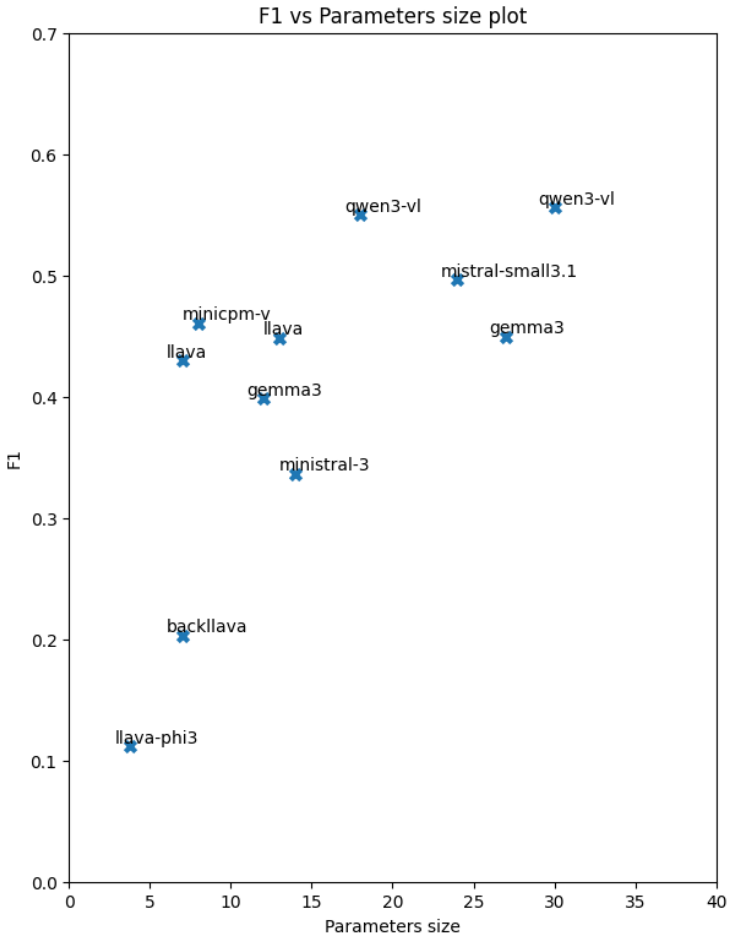
Zahŕňa pochopenie obsahu obrázka a jeho koreláciu s kontextom položenej otázky. Keďže musíme porovnať sémantiku informácií prítomných v oboch modalitách – v obrázku a v otázke v prirodzenom jazyku, ktorá s ním súvisí – VQA zahŕňa širokú škálu čiastkových problémov v CV aj NLP (ako je detekcia a rozpoznávanie objektov, klasifikácia scén, počítanie atď.). Preto sa považuje za úlohu kompletnú s využitím umelej inteligencie.

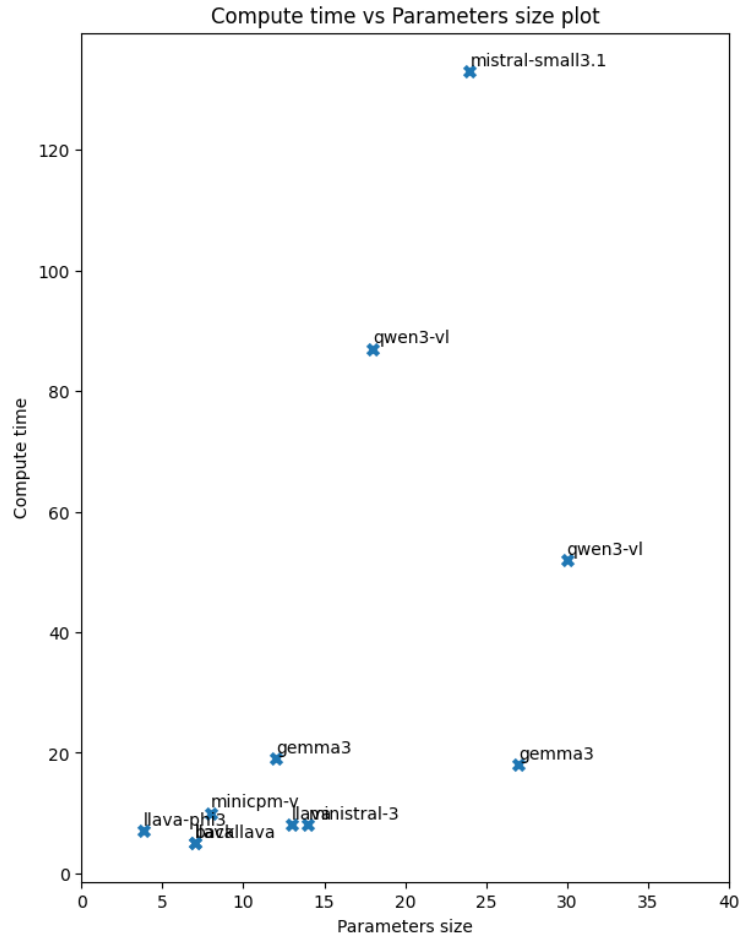
Precision and Reccal by BertScore





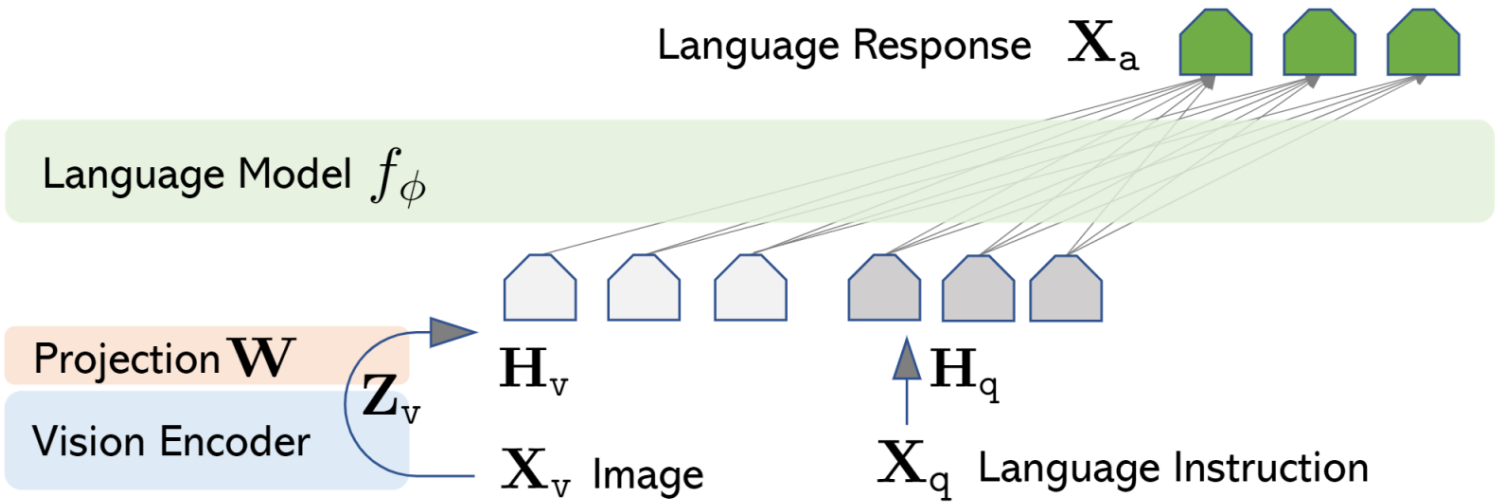
F1 by BertScore and Compute Time





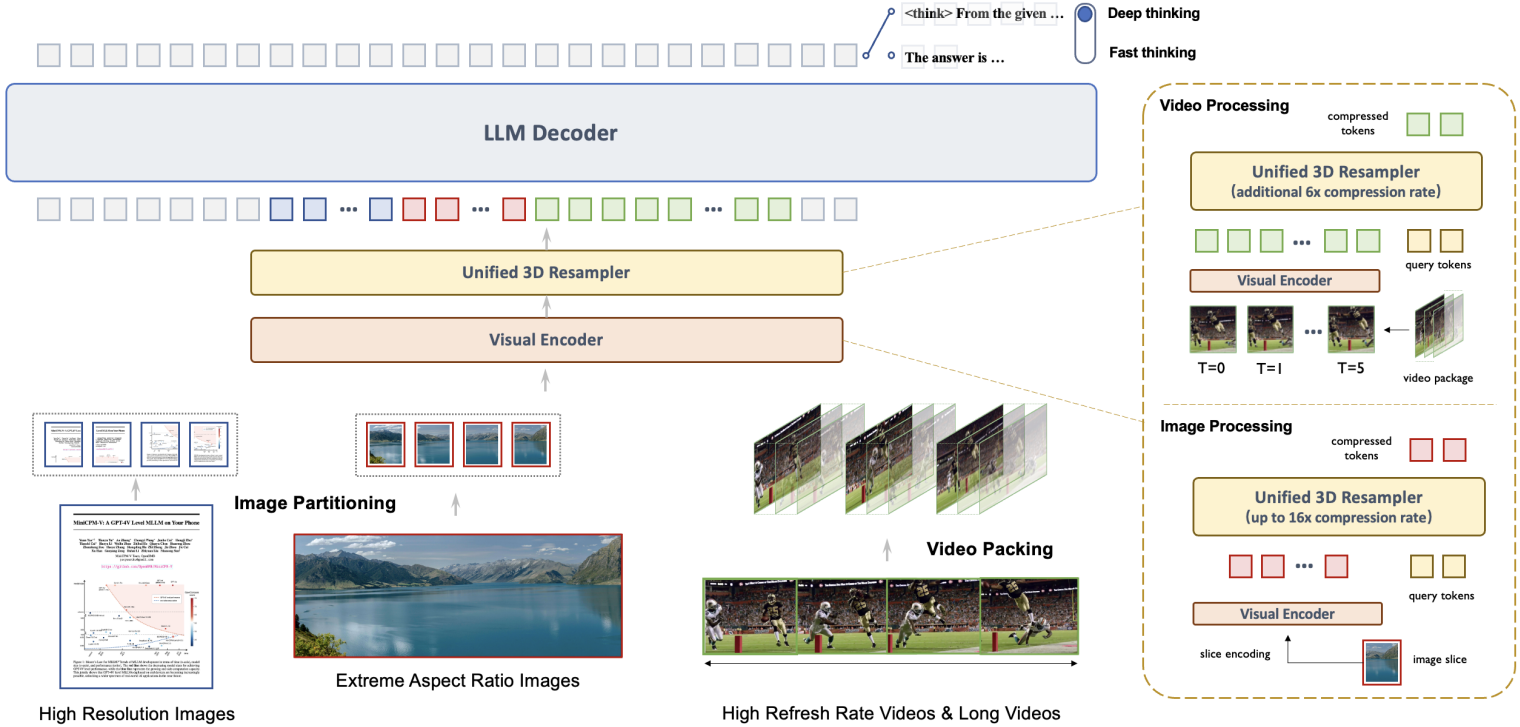
LLaVA Architecture

LLaVa spája vopred natrénovaný vizuálny enkodér CLIP ViT-L/14 a large language model Vicuna pomocou jednoduchej projekčnej matice.



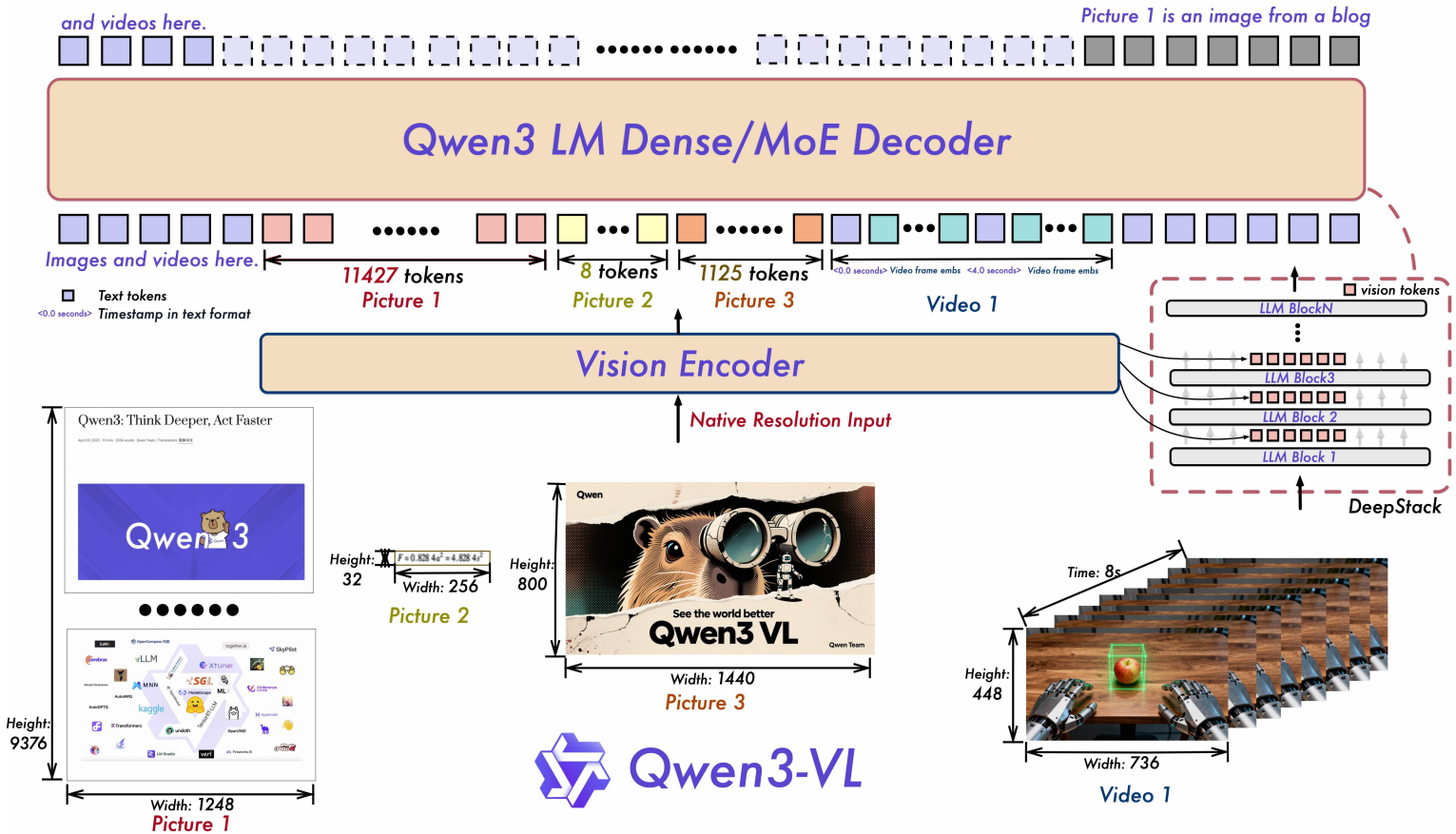
Minicpm-v Architecture

Model je postavený na modeloch SigLip-400M a Qwen2-7B s celkovým počtom parametrov 8B. Vykazuje výrazné zlepšenie výkonu oproti MiniCPM-Llama3-V 2.5 a prináša nové funkcie pre pochopenie viacerých obrazov a videa.



Qwen3-vl Architecture

- Interleaved-MRoPE
- DeepStack
- Zarovnanie textu a časovej pečiatky



Ďakujem za pozornosť