Co umí umělá inteligence vytěžit z dat webového archivu

Jan Lehečka, 18.10. 2022

Workshop
Data webových archivů a možnosti jejich využití

KATEDRA KYBERNETIKY

FAKULTA APLIKOVANÝCH VĚD
ZÁPADOČESKÉ UNIVERZITY
V PLZNI

LINDAT
CLARIAH-CZ

Osnova prezentace

- Trendy a architektura AI modelů
- Textové modely
- Audio modely
- Obrázkové modely

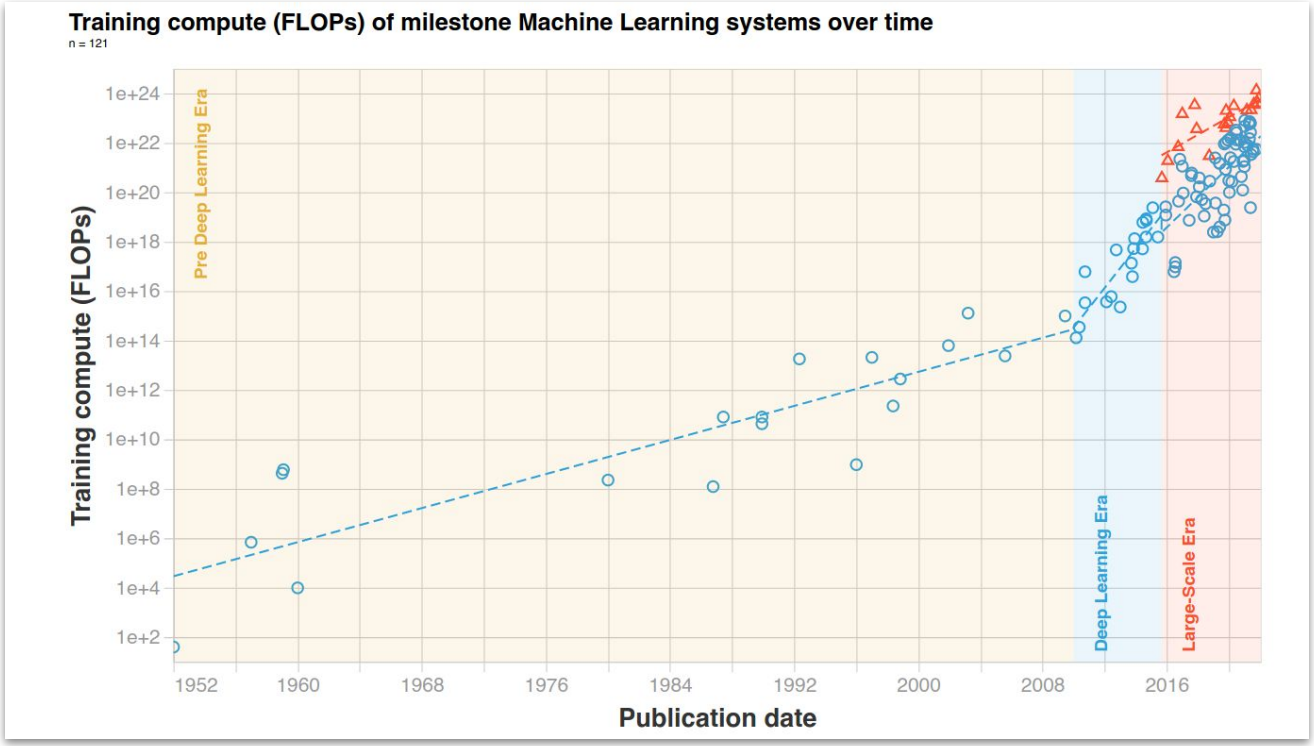
Trendy AI

rapidní zrychlení vývoje AI od r. 2010:

- ↑ velikost AI modelů
- ↑ velikost datasetů
- ↑ výpočetní náročnost trénování
- ↑ kvalita modelů

příčiny:

- masivní rozšíření internetu (data pro trénování modelů)
- nárůst výkonu HW (GPU)
- vývoj ženou dopředu technologičtí giganti (Google, Facebook, Microsoft, OpenAI, …)



Zdroj: Compute Trends Across Three Eras of Machine Learning (Sevilla et al., 2022)

Architektura současných AI modelů

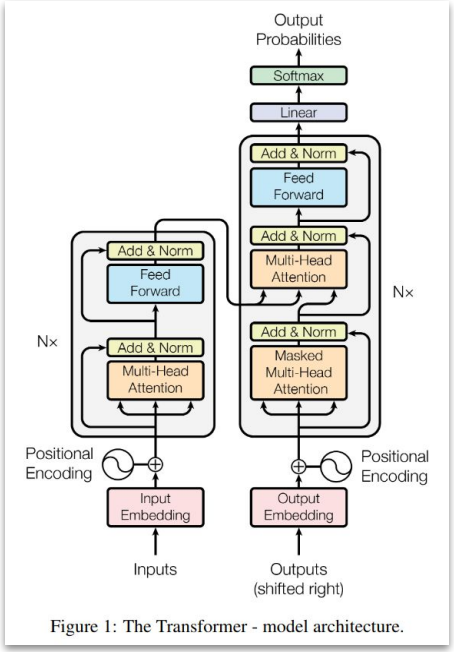
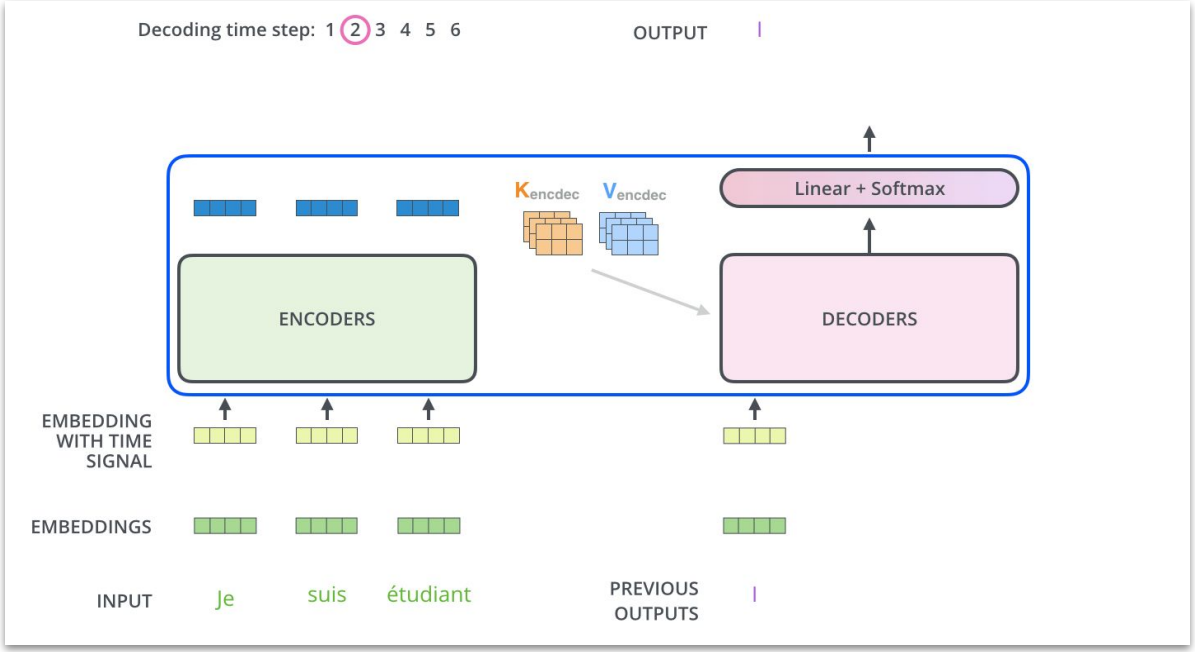


Figure 1: The Transformer - model architecture.

Zdroj: Vaswani, Ashish, et al. "Attention is all you need." 2017.



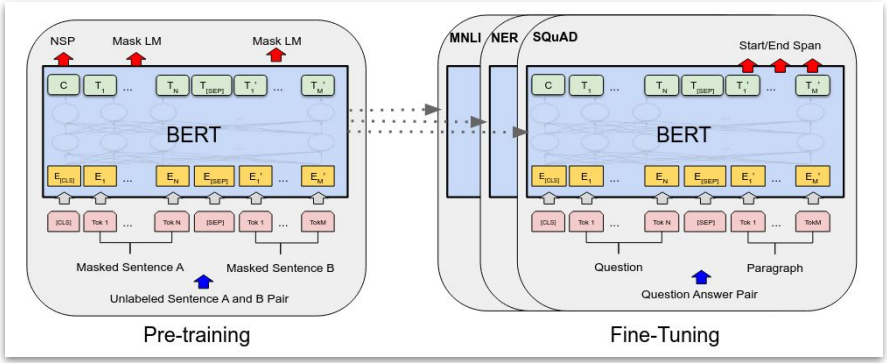
Zdroj: https://jalammar.github.io/illustrated-transformer

Textové modely – NLP, NLU, …

`[BERT, RoBERTa, DeBERTa, …]`

- před-trénink:
  - samoučení: doplňuje chybějící slova do vět
  - **je potřeba obrovské množství neanotovaných textů -> Webarchiv je ideální zdroj dat!**
- fine-tuning - vyžaduje anotovaná data
  - NLP - tokenizace, lemmatizace, POS, …
  - indexování & vyhledávání dokumentů
  - klasifikace/shlukování dokumentů podle:
    - témat
    - sentimentu
    - emocí
    - autora
    - …
  - NLU - strojový překlad, question answering, sumarizace, dialogové systémy, …

doba trénování (low-end GPU):
1 rok >> 1 den

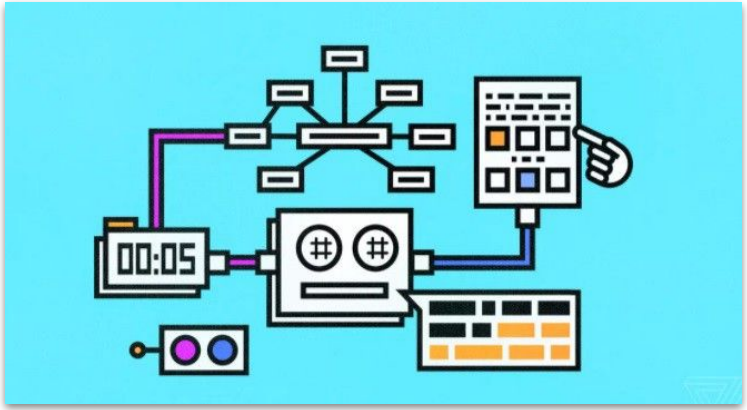


Zdroj: Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." (2018).

Textové modely – generátory textu, chatboty

`[GPT-2, GPT-3, T5, …]`

- modely, které generují text (článek, knihu, báseň, odpověď na dotaz, …)
- vhodným výběrem trénovacích dat lze ovlivnit “osobnost” bota:
  - konkrétní známá osobnost
  - vybraný autor (skupina autorů)
  - názorová skupina
  - návštěvníci určitého webu
  - “odborník” na nějaké téma
  - …

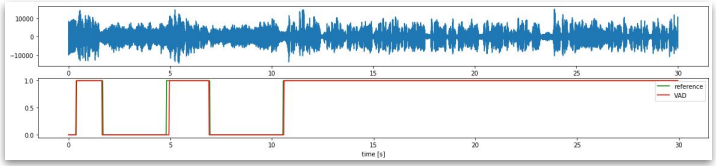


Zdroj: https://medium.com/swlh/everything-gpt-2-2-architecture-comprehensive-57129fac417a

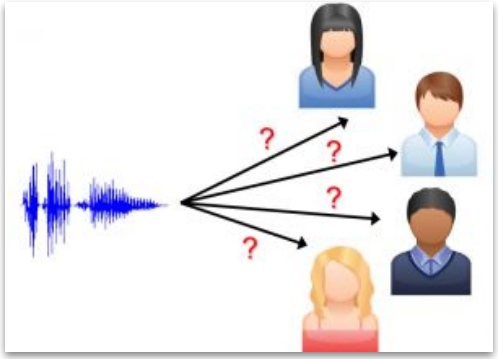
Audio modely

`[Wav2Vec2, HuBERT, Whisper, …]`

- před-trénink:
  - samoučení - učí se opravovat poškozený signál
  - **je potřeba obrovské množství neanotovaných nahrávek -> Webarchiv je ideální zdroj dat!**



- fine-tuning - vyžaduje anotovaná data
  - VAD - detekce, kde je v nahrávce řeč, hudba, ticho, …
  - identifikace jazyka, řečníka
  - ASR - rozpoznávání řeči
  - vyhledávání v audiu (keyword spotting)
  - audio-to-audio - odstranění šumu, rozdělení překrývajících se promluv
  - v kombinaci s textovým vstupem:
    - TTS - syntéza řeči určitého řečníka

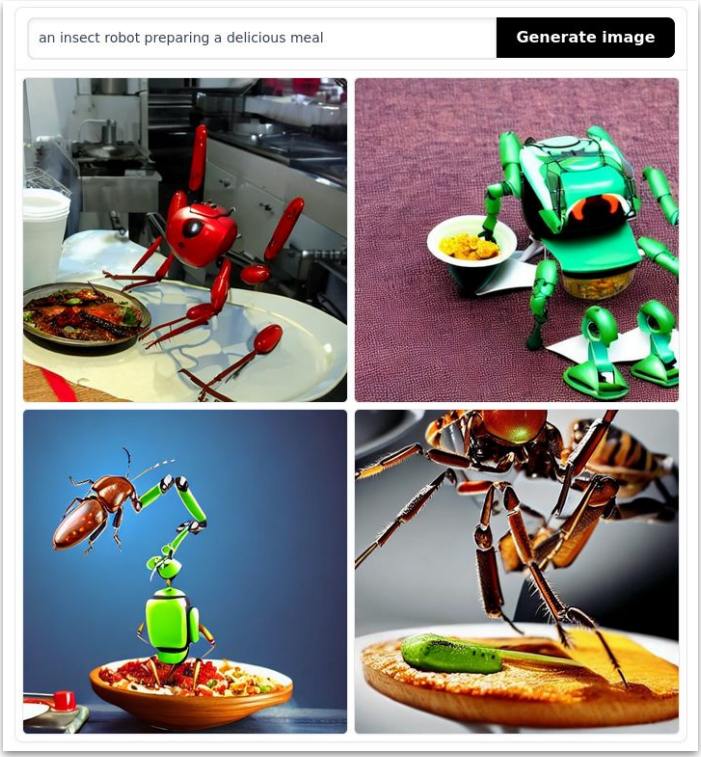


Zdroj: https://partner.phonexia.com/wp-content/uploads/2018/03/speaker_identification-300x215.png

Obrázkové modely

[DALL·E 2, Stable Diffusion, …]

- **Webarchiv obsahuje velké množství obrázků => je ideální zdroj dat pro před-trénink!**
- fine-tuning - vyžaduje anotovaná data
  - klasifikace / shlukování obrázků
  - segmentace obrazu
  - detekce objektů
  - popis obrázku
  - v kombinaci s textovým vstupem:
    - generování obrázků
    - generování videí



Zdroj: https://huggingface.co/spaces/stabilityai/stable-diffusion

Závěr

- state-of-the-Art AI modely vyžadují obrovské množství trénovacích dat bez anotací

- Webarchiv:
  - ideální zdroj dat pro před-trénink AI modelů
  - AI modely lze použít pro zpracování dat Webarchivu
    - přiřazení metadat k záznamům (text, audio, obrázky)
  - obohacení Webarchivu o vytěžená metadata umožňuje velmi pokročilé vyhledávání a analýzu dat pro široké spektrum vědeckých disciplin

Děkuji za pozornost