Umělá inteligence a její využití v programování

KIV/ADT – 12. přednáška

Ing. Miloslav Konopík, Ph.D.
konopik@kiv.zcu.cz

10. května 2024

Obsah

Jazykové modely

Chatboti

Chatbot

- Počítačový program, který simuluje lidskou konverzaci.
- Může být použit pro odpovídání na otázky, poskytování informací, zpracování příkazů nebo zábavu.
- Založený na umělé inteligenci a strojovém učení.
- Využívá **jazykové modely** pro generování textů.

Chatbot

- Odpovídá na otázky a poskytuje informace.
  - Pamatuje si historii a reaguje v kontextu.
  - Výpočetně náročný.

Vyhledávač

- Hledá informace na základě klíčových slov.
  - Nezohledňuje kontext.
  - Velmi rychlý a efektivní.

Jazykové modely

Hlavní úkol jazykových modelů:

- Odhadnout pravděpodobnost sekvence slov (výskytu sekvence náhodné proměnné):

$$P(w_1^k) = ? \tag{1}$$

Lze použít k předpovědi následujícího slova:

$$P\left(w_i | w_1^{i-1}\right) = P\left(w_i | w_{i-1}, w_{i-2}, \ldots, w_1\right) \tag{2}$$

Například:

V Plzni se narodil ...

Jazykové modely

Hlavní úkol jazykových modelů:

- Odhadnout pravděpodobnost sekvence slov (výskytu sekvence náhodné proměnné):

$$P(w_1^k) = ? \tag{1}$$

Lze použít k předpovědi následujícího slova:

$$P\left(w_i | w_1^{i-1}\right) = P\left(w_i | w_{i-1}, w_{i-2}, \ldots, w_1\right) \tag{2}$$

Například:

V Plzni se narodil Karel Gott.

Jazykové modely

Hlavní úkol jazykových modelů:

- Odhadnout pravděpodobnost sekvence slov (výskytu sekvence náhodné proměnné):

$$P(w_1^k) = ? \tag{1}$$

Lze použít k předpovědi následujícího slova:

$$P\left(w_i | w_1^{i-1}\right) = P\left(w_i | w_{i-1}, w_{i-2}, \ldots, w_1\right) \tag{2}$$

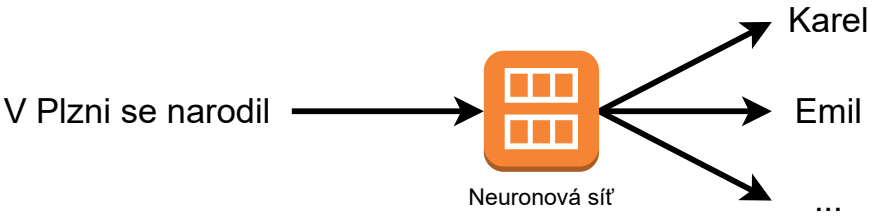
Například:

V Plzni se narodil Emil Škoda.

N-gramové modely

Historie	Následující slovo	Počet
v Plzni	byla	1 161
	působí	763
	se	354
Plzni se	ceny	342
	rozhodlo	131
	narodil	127
se narodil	v	1 769
	jako	1 213
	Karel	67
	Emil	12

Použití neuronových sítí



Jazykové modely – učení

Strategie učení:

- Předpovídání schovaného slova.
- Odšumování (opravování, hledání správného pořadí, ...).
- Generování pokračování – vyžaduje dekodér.

Jazykové modely – učení

Strategie učení:

- Předpovídání schovaného slova.
- Odšumování (opravování, hledání správného pořadí, ...).
- Generování pokračování – vyžaduje dekodér.

Dekodér:

- Algoritmus pro generování optimální sekvence pokračování.
- Paprskové prohledávání (beam search).

Jazykové modely – učení

Strategie učení:

- Předpovídání schovaného slova.
- Odšumování (opravování, hledání správného pořadí, ...).
- Generování pokračování – vyžaduje dekodér.

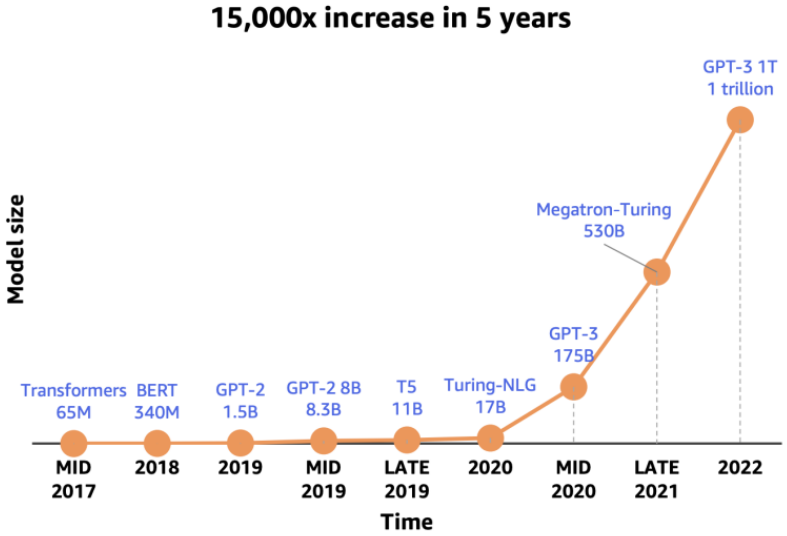
Dekodér:

- Algoritmus pro generování optimální sekvence pokračování.
- Paprskové prohledávání (beam search).

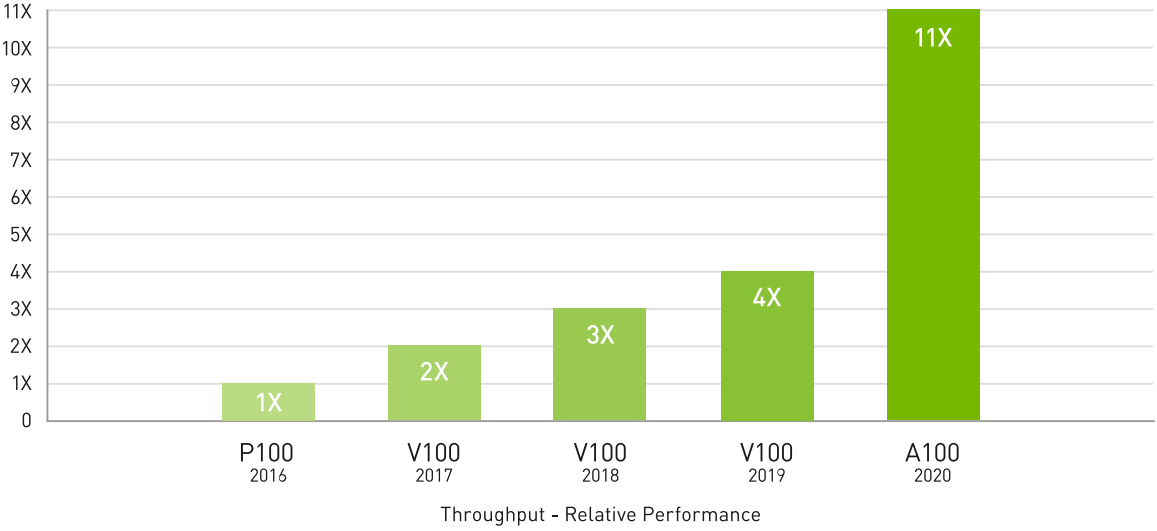
Společná vlastnost učení jazykových modelů:

- Stačí čistý (neoznačený) text.

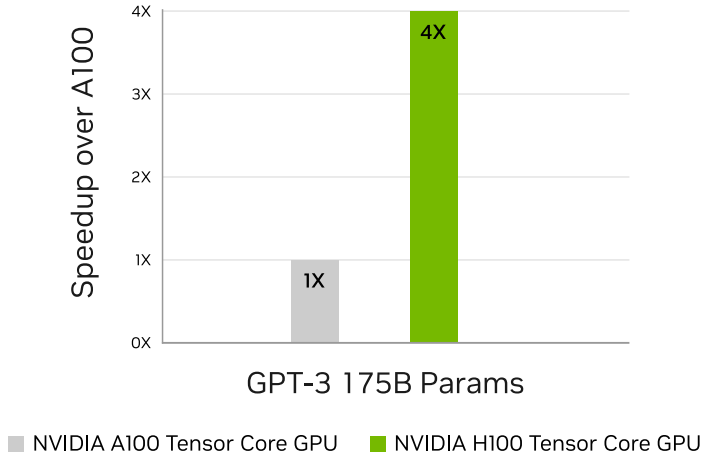
Velké jazykové modely



Hardware pro trénování



Hardware pro trénování



Zpracování textu v počítači

Uložení textu v počítači

Text je v počítači uložen v číslech:

Český novinář Ferdinand Peroutka kdysi podotkl: ...

225	1225	25226	25227	857	625	...

n	o	v	i	n	á	ř
110	111	118	105	110	225	354

Uložení textu v počítači

Text je v počítači uložen v číslech:

Český novinář Ferdinand Peroutka kdysi podotkl: ...

225	1225	25226	25227	857	625	...

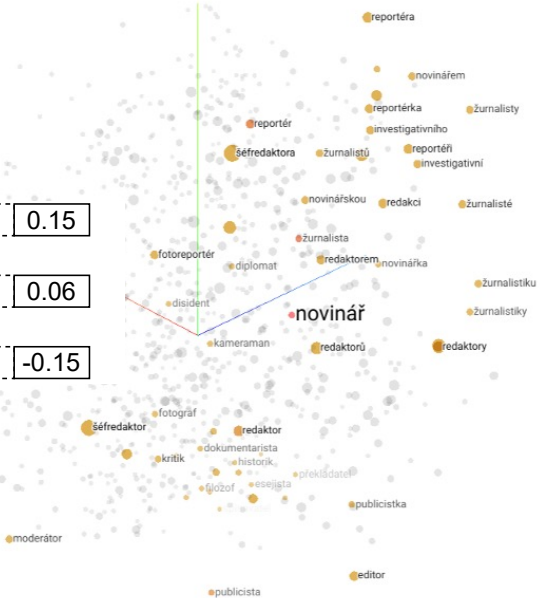
n	o	v	i	n	á	ř
110	111	118	105	110	225	354

Novinář:

- publicista: 77351
- žurnalista: 15547
- komentátor: 35547
- spisovatel: 752
- sporťák: 11567
- osoba: 463

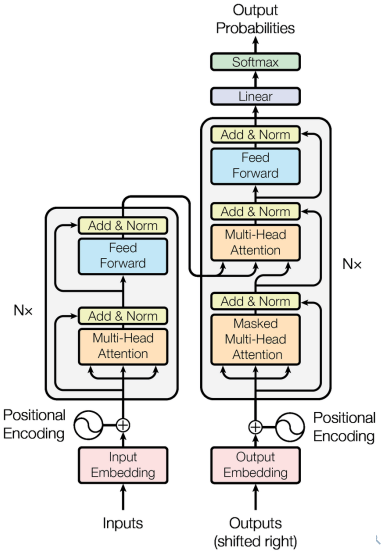
Slovní vektory

novinář	→	1225	→	-0.27	0.23	...	0.15
publicista	→	77351	→	-0.18	0.22	...	0.06
okno	→	3351	→	-0.30	0.07	...	-0.15

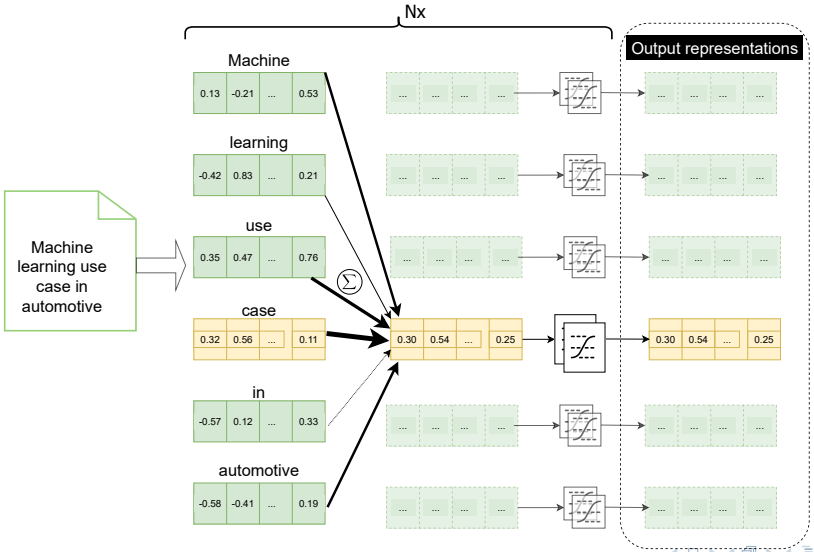


Transformer – Model

- Umožňuje dávat slova do souvislostí a vytvořit číselné reprezentace významu vět.
- Používá mechanismus pozornosti.
- Základní mechanismus pozornosti vyžaduje kvadratickou výpočetní náročnost vzhledem k délce vstupu.



Transformer – příklad



Dotazování modelů.

- **Dotaz** – text $\vec{x'}$ vytvořený ze vstupního dotazu $\vec{x}$ pomocí *šablony*
- Šablona – text s dvěma sloty: vstupním slotem `[X]` pro vstup $\vec{x}$ a slotem `[Z]` pro odpověď $\vec{z} \rightarrow$ mapováno na $\vec{y}$

- Příklad pro odpovídání otázek:
  - $\vec{x'}$: „Kdo se narodil v Plzni“
  - Šablona: V `[X]` se narodil `[Z]`.
  - $\vec{x}$: „V Plzni se narodil“
  - Odpověď: „V Plzni se narodil Karel Gott.“

Datové sady dialogů.

Modely jsou dále doladěny na datových sadách dialogů.
Struktura:

- Vstup: otázka.
- Výstup: očekávaná odpověď
- Volitelně:
  - Reakce na odpověď.
  - Doplňující odpověď.
- Metadata:
  - Kvalita
  - Kreativita
  - Humor
  - Slušnost
  - ...

Datové sady dialogů.

Modely jsou dále doladěny na datových sadách dialogů.
Struktura:

- Vstup: otázka.
- Výstup: očekávaná odpověď
- Volitelně:
  - Reakce na odpověď.
  - Doplňující odpověď.
- Metadata:
  - Kvalita
  - Kreativita
  - Humor
  - Slušnost
  - ...

Lidská data

Nejcennější je živá zpětná vazba od lidí. Hodnotí aktuální průběh dialogu.

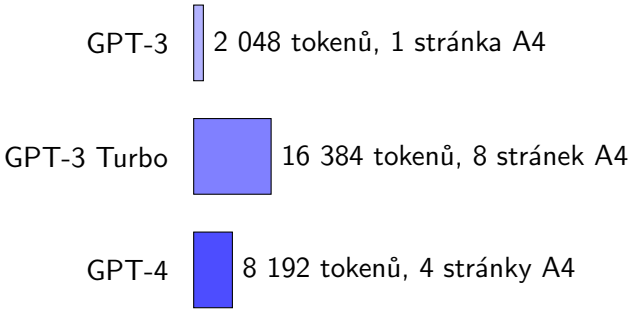
Řízení dialogu.

- Systémové dotazy (skryté, modifikují chování modelu).
- Kontrola výstupu (jiný AI / LM, který cenzuruje výstupy).
- Omezení délky dialogu (kvadratická výpočetní náročenost vzhledem k délce vstupu – omezení délky kontextu).

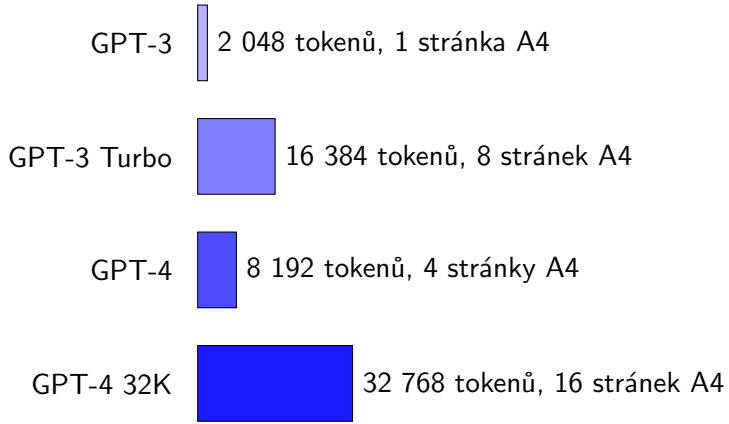
Další poznámky.

- Cena za provoz – 36 centů (přibližně 10x více než vyhledávání klíčovými slovy).
- Open source alternativy.
  - Open Assistant.
  - GPT for All
  - OpenChatKit
- Další vylepšení:
  - Agenti.
  - Myšlenkový řetěz – vylepšování dotazů.
  - Plug-iny, nástroje.
  - AutoGPT (automatické dotazování + nástroje).

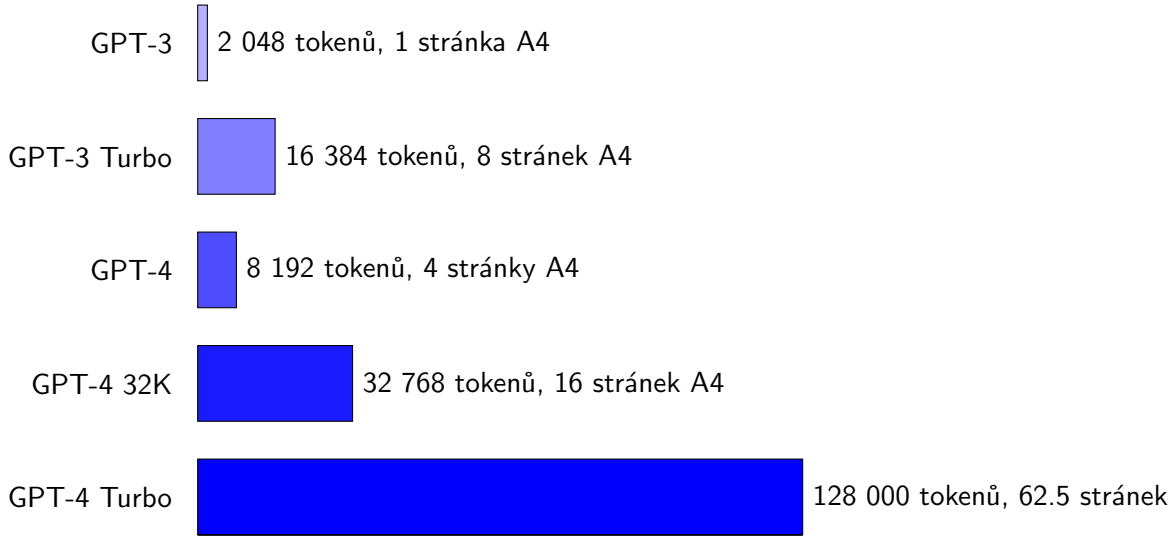
Pokrok velkých jazykových modelů s rostoucí délkou kontextu 2023



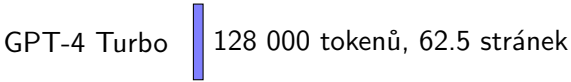
Pokrok velkých jazykových modelů s rostoucí délkou kontextu 2023



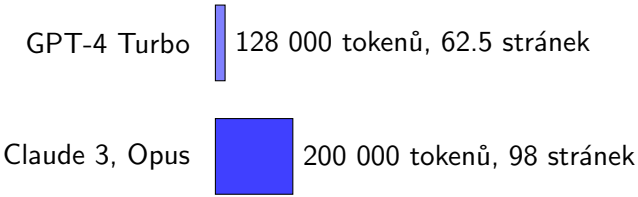
Pokrok velkých jazykových modelů s rostoucí délkou kontextu 2023



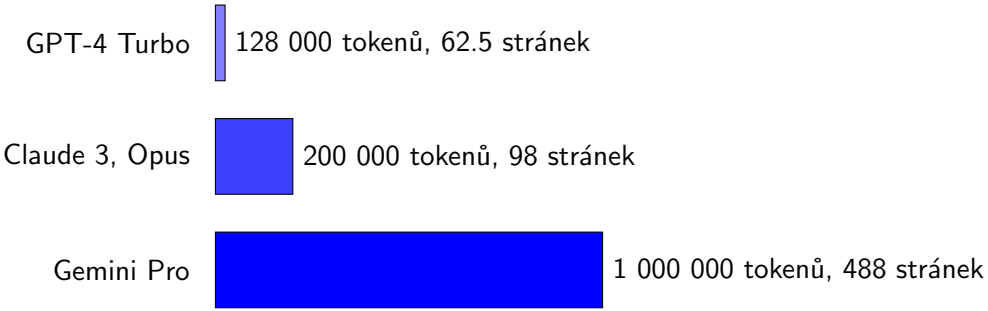
Pokrok velkých jazykových modelů s rostoucí délkou kontextu 2024



Pokrok velkých jazykových modelů s rostoucí délkou kontextu 2024



Pokrok velkých jazykových modelů s rostoucí délkou kontextu 2024



Využití v programování

Využití v programování

- Generování kódu.
- Dokumentace.
- Testování.
- Vývojové prostředí.
- ...

Praktické ukázky

- ChatGPT plus.
- Github Copilot.

Vyhodnocení předmětu

Dotazník

References I