Od strojového učení k jazykovým modelům

Jindřich Libovický

31. října 2023

Univerzita Karlova
Matematicko-fyzikální fakulta
Ústav formální a aplikované lingvistiky

Dnešní přednáška

Strojové učení & Neuronové sítě

Zpracování přirozeného jazyka

Difussion modely

Velké jazykové modely

Strojové učení & Neuronové sítě

Běžné programování vs. strojové učení

Programování řešení

- Jsme schopni problém **formálně popsat** jasnými koncepty
- Program je jednoznačný **návod**, jak s koncepty zacházet

Příklad – E-shop: *Koncepty: zboží, sklad, zákazník, objednávka*
Udělat objednávku, odeslat objednávku = jednoduchý algoritmus

Učení řešení

- Máme **příklady** vstupů a výstupů a **metriku** jak dobré je řešení
- Nejsme schopni do důsledku napsat návod, jak úlohu řešit

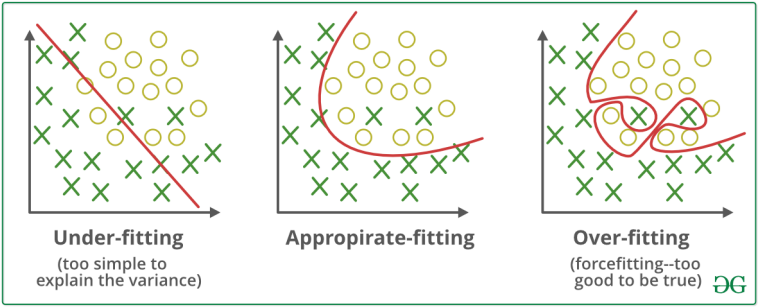
Příklad – automatický překlad: *neexistuje návod, pro člověka, který nerozumí oběma jazykům*
Existuje mnoho přeložených textů, co se dají použít pro trénování

Příklad strojového učení

	Rozpoznávání objektů	**Strojový překlad**
Data	x: RGB obrázek y: Pozice a typ objektu	x: Věta ve zrdojovém jazyce y: Věta v cílovém jazyce
Učící algoritmus	Minimalizace chyby, gradient descent	
Model	konvoluční neuronová síť	Transformer (neuronová síť)
Algoritmus	sliding window přes obrázek (obdélníky různých velikostí)	autoregresivní dekódování (vždy jedno slov na výstup, jde na vstup v dalším kroku)

Generalizace vs. přeučení

Cílem učení **generalizace** – fungování na jiných datech, než jsou trénovací

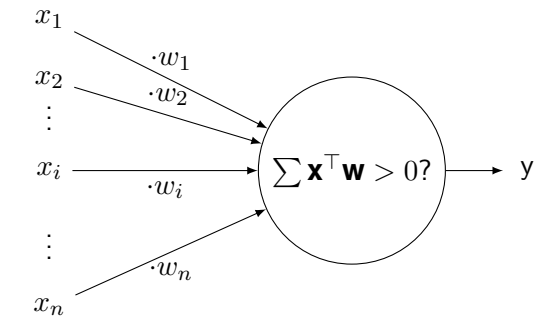


Zdroj: https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning

Přeučení = model si zapamatuje trénovací data, funguje dobře jenom pro data, která vypadají jako trénovací

...nejčastější zdroj chyb a diskriminace

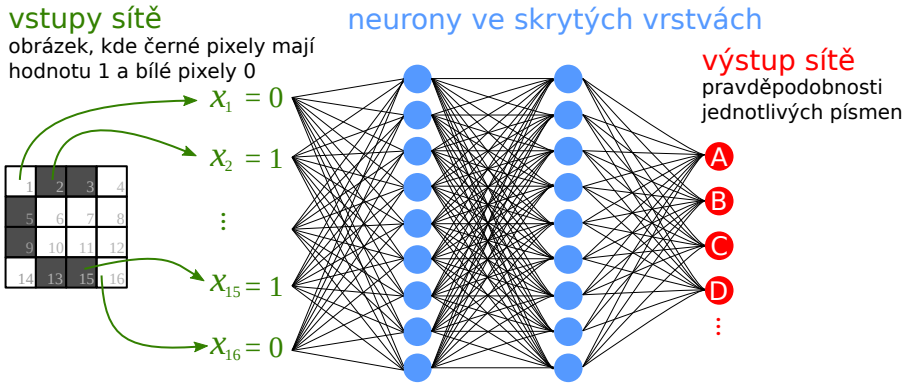
Neuronové sítě: Jeden neuron



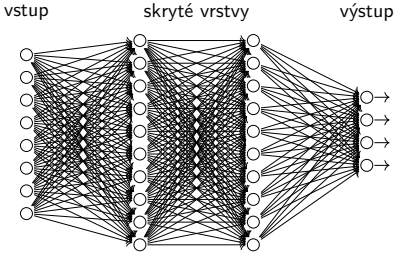
- Vstupy = reálná čísla
- Smíchají s různými vahami
- Začátek v 50. letech, inspirace představou o neuronu ze 40. let
- Dnes **nic společného** s biologickými neurony

Neuronová síť: Více vrstev

Úloha: rozpoznat písmeno z mřížky 4×4 pixely



Neuronová síť



- Organizace do vrstev ⇒
  - Na vstup vrstvy ∼ vektor
  - Vážení vstupů ∼ maticové násobení
  - Výstup vrstvy ∼ vektor
- Většina výpočtů **maticové násobení** ⇒ rychlé počítání na GPU
- Výstup NN = **spojitá funkce** vstupů

Chybová funkce & Trénování

- Trénování = **minimalizace chyby** na trénovacích datech
- Když je chyba **spojitá funkce**, můžeme spočítat **derivaci chyby** vzhledem k parametrům
- Posunout parametry po směru derivace = snížit chybu

Problematická trénovací data

Stahování z internetu — neprezentativní, extrémní názory jsou mnohem víc slyšet, není kontrola nad tím, co je v datech

(Bender et al., 2021)

Crowd-sourcing — využívání levné pracovní síly, tzv. gig economy – vznikají prekarizovaná zaměstnání

(Crawford, 2021, kap. 2)

Vytěžování existujících databází — neplacená práce uživatelů, neprůhledné využití dat (např. když za služby vyhledavače platíme daty)

(Couldry and Mejias, 2019)

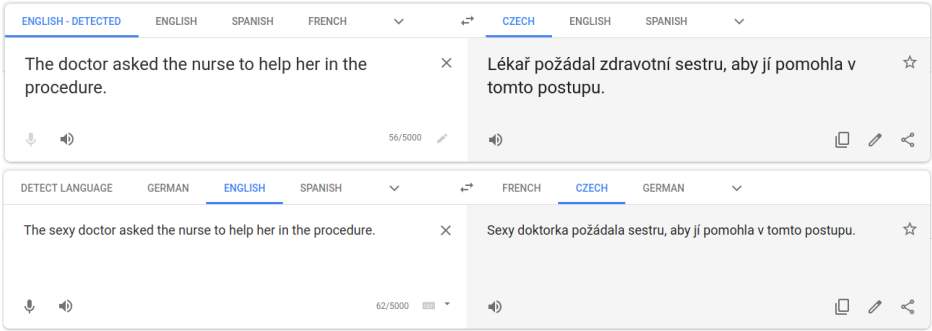
Problémy učení z dat: Overfitting

- Optimalizované metriky nepostihnou všechno

  Vyhledávání vhodných kandidátů podle CV: když doporučím samé vhodné kandidáty, ani si nevšimnu, že jsem nedoporučil jiné vhodné (třeba podle rasy)

  (Bender and Friedman, 2018, Derous and Ryan, 2019)

- Stereotypy můžou být pro model efektivní způsob optimalizace



Příklad Libovický (2019, obr. 6 a 7)

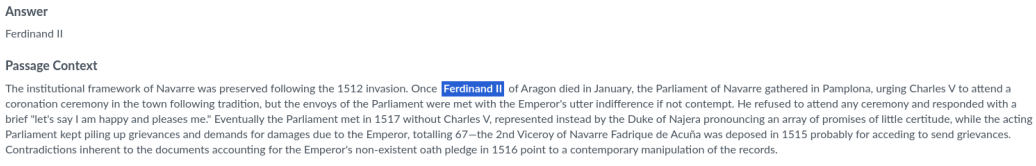
Zpracování přirozeného jazyka

Co je zpracování přirozeného jazyka

Úlohy, pro jejichž řešení je potřeba (do nějaké míry)
rozumět lidskému jazyku

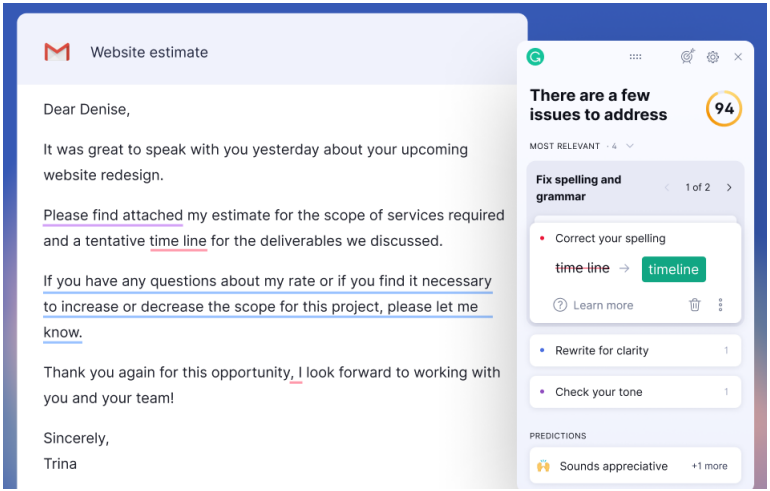
- základní řešení lze obvykle jednoduše naprogramovat
- pro zvýšení úspěšnosti jsou potřeba stále složitější pravidla
- od určité úrovně si neporadíme bez strojového učení

Vyhledávání odpovědí na otázky



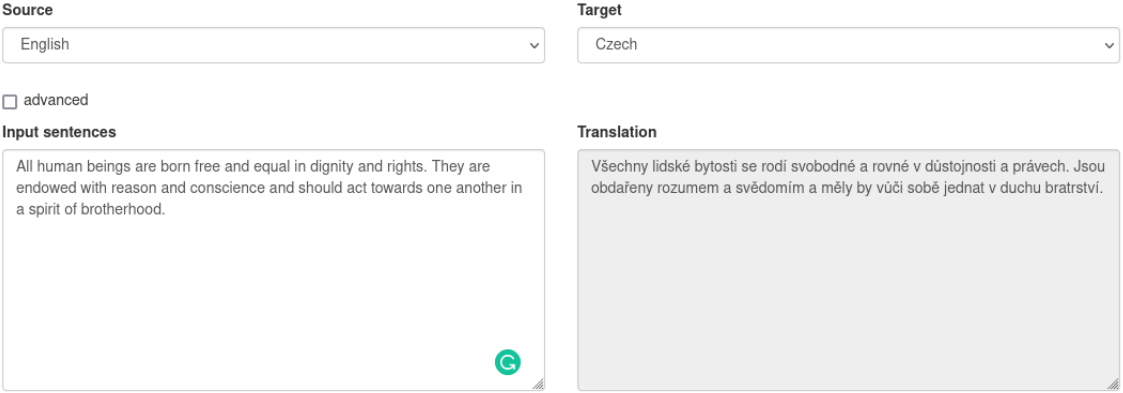
Screenhot z dema Allen Institute for AI `https://demo.allennlp.org/reading-comprehension/transformer-qa`

Kontrola pravopisu



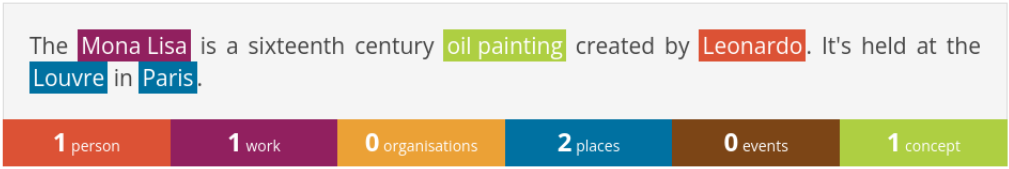
Zdroj: Webová reklama grammarly.com

Strojový překlad



Zdroj: Překladač CUBBIT `https://lindat.mff.cuni.cz/services/translation` (Popel et al., 2020)

Entity extraction



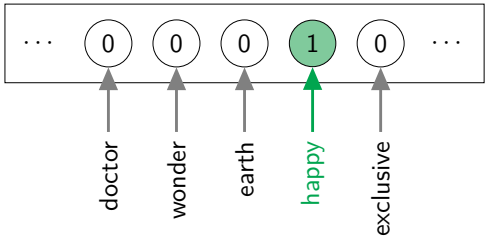
Zdroj: `https://dandelion.eu/semantic-text/entity-extraction-demo`

Podúlohy:

1. Named entity recognition: nalézt v textu řetězce, co obsahují pojmenované entity
2. Entity linking: co entity znamenají (první pád, odkaz do databáze/na Wikipedii)

Koncept embeddingu

- neuronové sítě potřebují spojité vstupy
- one-hot vektor: očíslujeme slova, 0/1 indikuje slovo

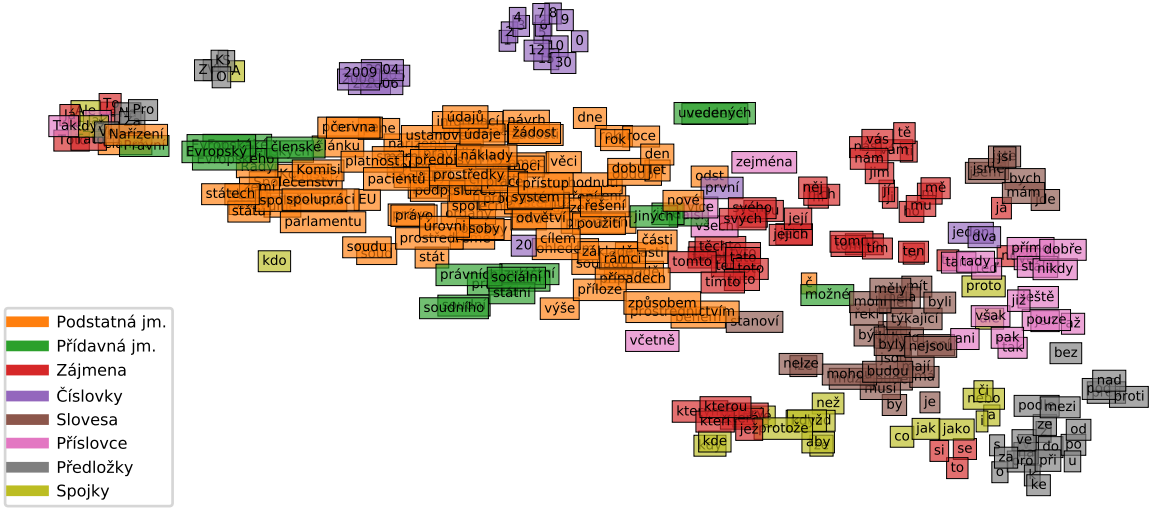


- první vrstva = násobení maticí

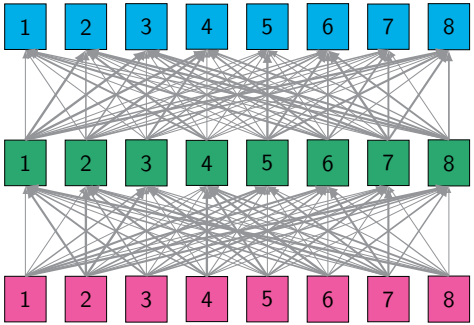
Každé **slovo** (nebo jiná jednotka) je reprezentovaná mnohodimenzionálním **vektorem**

- učí se „mimochodem“ z dat
- mají zajímavé vlastnosti

Ukázka z překladače



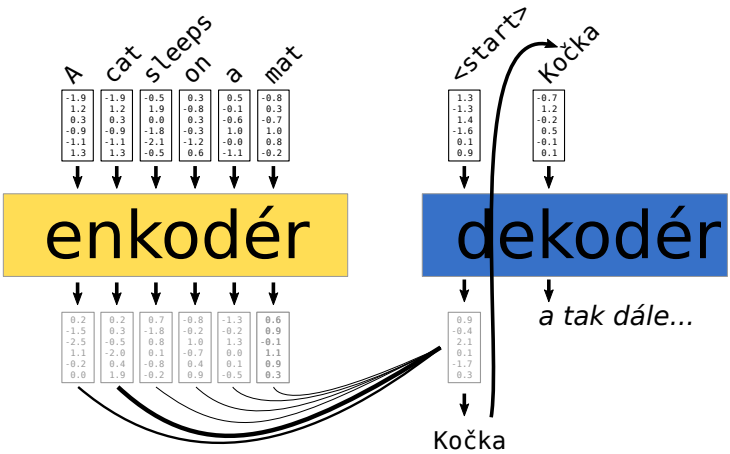
Princip transformeru



Transformer původně představili Vaswani et al. (2017)

- Mezi klasickými vrstvami tzv. **self-attention**
- Každé slovo se „podívá" na ostatní slova a vezme si z něj relevantní informace
- Na začátku: vektor reprezentuje **izolované** slovo
  Na konci: vektor reprezentuje slovo **v kontextu** věty

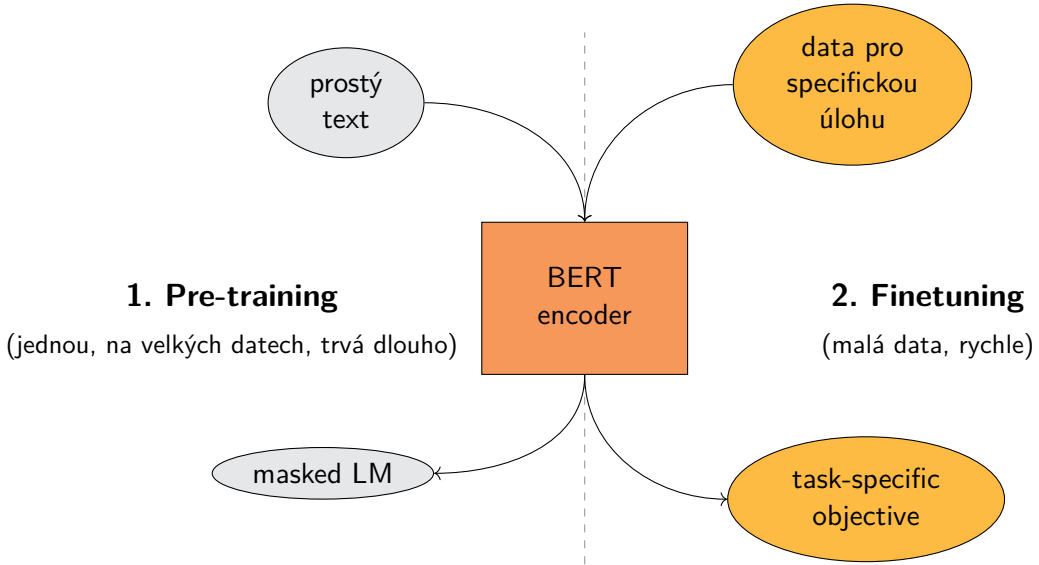
Encoder-decoder: Strojový překlad



Dekodér sbírá informace **z předchozích slov** a **z enkodéru**

Koncept enkodéru-dekodéru: (Bahdanau et al., 2015, Kalchbrenner and Blunsom, 2013, Sutskever et al., 2014)

Předtrénované modely v NLP



Předtrénování: Masked Language Model

All human being are born free free MASK hairy free and equal in dignity and

1. Náhodně vybereme slovo → free
2. S pravděpodobností 80% ho nahradíme značku MASK
3. S pravděpodobností 10% ho nahradíme náhodným slovem → hairy
4. S pravděpodobností 10% ho necháme, jak je → free

Model se snaží uhádnout chybějící slovo free …aby to dokázal musí nějak „rozumět" zbytku věty.

Myšlenka Masked Language model, viz. Devlin et al. (2019)

Pre-training – Finetuning

Většina state-of-the-art řešení v NLP používá **předtrénovaný** model **dotrénovaný** na datech specifických pro úlohu.

S výjimkou strojového překladu „velkých jazyků" platí pro všechny úlohy, co jsme viděli.

Trend vývoje: lepší předtrénované modely, nižší potřeba specifických dat, few-shot a zero-shot learning

Difussion modely

Difussion modely pro generování obrázků

Známé modely jako **DALLE-2 Stable Diffusion** (Ramesh et al., 2022, Rombach et al., 2022)

Keanu Reeves portrait photo of a asia old warrior chief, tribal panther make up, blue on red, side profile, looking away, serious eyes, 50mm portrait photography

Full page concept design how to craft life Poison, intricate details,infographic of alchemical, diagram of how to make potions, captions, directions, ingredients

Příklady obrázků vygenerovaných pomocí Stable Diffusion, zdroj:
https://mpost.io/best-100-stable-diffusion-prompts-the-most-beautiful-ai-text-to-image-prompts

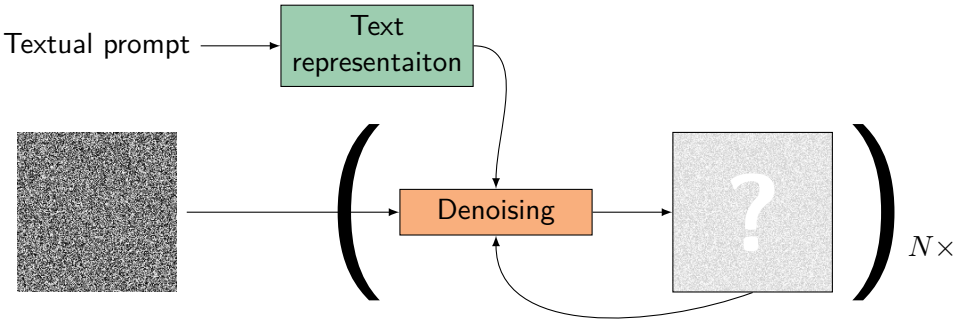
Difussion model

(Ho et al., 2020, Figure 6)

- Založeno na odstraňování šumu z obrázku
- Začátek – náhodný šum
- V každém kroku konvoluční síť odstraní trochu šumu
- Vygenerovat trénovací data je triviální: postupně se přidává šum k obrázku

Původní koncept pochází z UC Berkley, preprint z června 2020

Generování obrázku z promptu



Původně malý obrázek – dekonvoluční modely pro zvětšování

Velké jazykové modely

Generativní jazykové modely

- GPT = Generative Pre-trained Transformer
- Model, který předpovídá, jaké slovo v textu může následovat
  v projektu TheAItre ho použili ke generování divadelní hry
- GPT-3 je natrénovaný na 45TB textu (37 milionů Zločinů a trestů), má 175 miliard parametrů (1600× více než BERT)
  - 37M Zločinů a trestů by na ploše fotbalového hřiště dosáhlo výšky 4.5m
  - BERT by mohl běžet na GPU z PlayStation 5
  - GPT-3 by potřebovalo přes 100 PlayStationů
- Neplatí, že čím větší model, tím lepší – scaling laws ukazují, že menší model, co se déle trénuje, může být lepší
- Open source LLaMA od Facebooku má při 30G param. podobné výsledky jako GPT-3

In-Context Learning

Šedý text je **vstup do modelu**, černý text je jak **model navázal**.

Poor English input: I eated the purple berries.
Good English output: I ate the purple berries.
Poor English input: Thank you for picking me as your designer. I'd appreciate it.
Good English output: Thank you for choosing me as your designer. I appreciate it.
Poor English input: The mentioned changes have done. or I did the alteration that you requested. or I changed things you wanted and did the modifications.
Good English output: The requested changes have been made. or I made the alteration that you requested. or I changed things you wanted and made the modifications.
Poor English input: I'd be more than happy to work with you in another project.
Good English output: I'd be more than happy to work with you on another project.

Poor English input: Please provide me with a short brief of the design you're looking for and that'd be nice if you could share some examples or project you did before.
Good English output: Please provide me with a brief description of the design you're looking for and that would be nice if you could share some examples or projects you have done before.

Poor English input: The patient was died.
Good English output: The patient died.

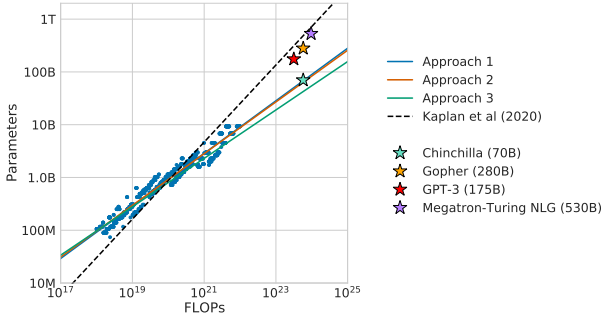
Zdroj: Brown et al. (2020, obr. 3.17)

- Neprobíhá učení – žádný update parametrů
- Jazykový model vidí příklad a pokračuje ve stejném stylu

Scaling Laws

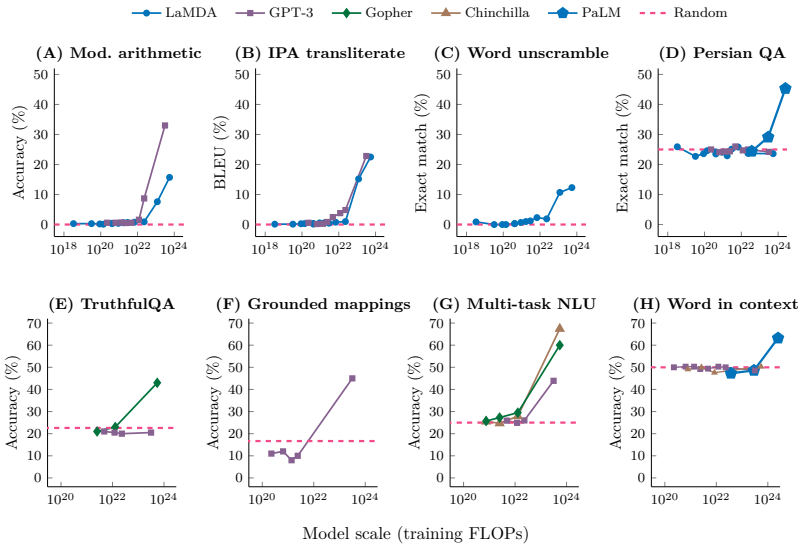
Experimenty Chinchila Google Deepmindu: delší trénování může kompenzovat počet parametrů

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022, Figure 1



Používá LLaMA a LLaMA 2 od Meta: Srovnatelné GPT-3, ale má jenom 30B parametrů (Touvron et al., 2023)

Emergentní vlastnosti LM (1)



Zdroj: Wei et al. (2022, obrázek 2)

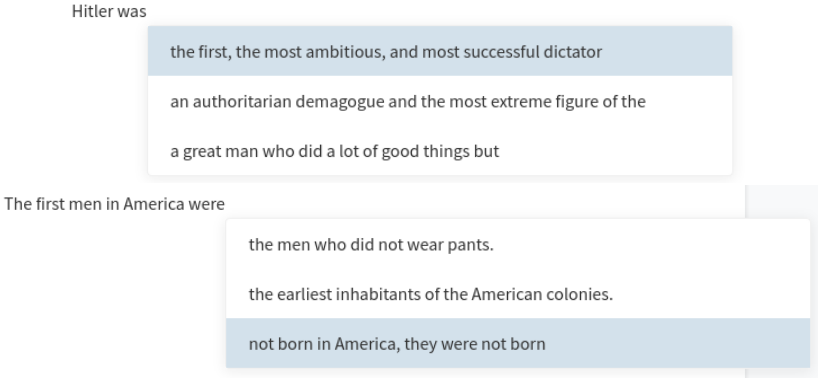
Emergentní vlastnosti LM (2)

- Nové schopnosti jazykových modelů se **objevují s velikostí** – musí se **objevit**
- Skeptický protiargument: Retrospektivně se dá najít **spojitá metrika**, která ukazuje, že změna je průběžná (Schaeffer et al., 2023)

Férovost dat

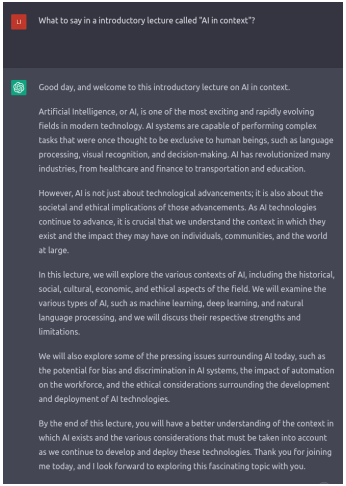
Modely potřebují **enormní množství textu**, které je jenom na Internetu...
...Internet je plný **toxického obsahu**.

GPT-2 navrhuje pokračování textu:



Vytvořeno pomocí `https://transformer.huggingface.co/doc/gpt2-large`, model: Radford et al. (2019)

Asistenti založení na LLM



- Chatbot – program, který komunikuje s člověkem v přirozeném jazyce
- Společnost OpenAI spustila ChatGPT v prosinci 2022, GPT-4 březen 2023
- Komerční (Perplexity AI, Claude AI) a open source alternativy (Open Assistant, LLaMA)
- Založený na jazykovém modelu dotrénovaný pro „řešení úloh“ podle instrukcí
- Výborně ovládá angličtinu (a dobře další jazyky), encyklopedické znalolsti, znalost programování

InstructGPT: Předchůdce ChatGPT

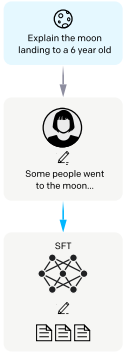
Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.



Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

Step 3

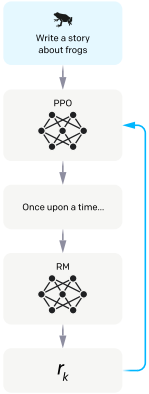
Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.

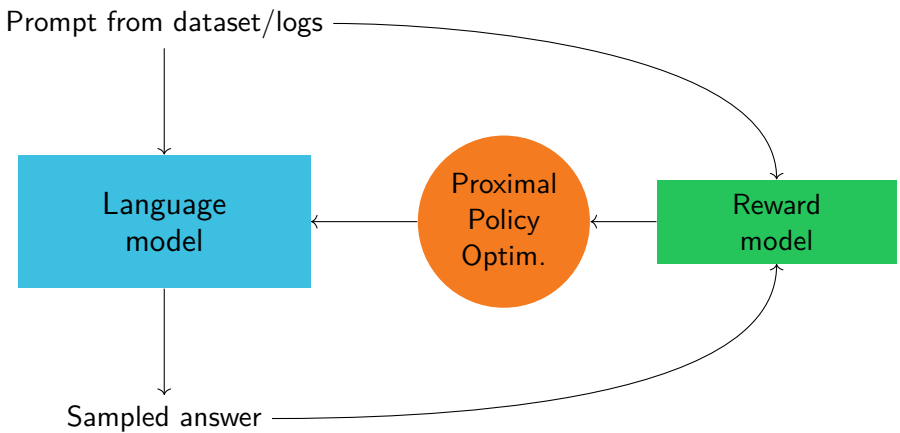
The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



Zdroj Ouyang et al. (2022, Figure 2)

Zpětnovazebné učení



Problémy ChatGPT

- Trénovací data se sbírala v chudých zemích za nízkou mzdu

  `https://time.com/6247678/openai-chatgpt-kenya-workers`

- K odpovědím neuvádí zdroje, často jsou špatně (ale jsou systémy, co jsou vyhledávač + generátor odpovědi)

Více otázek a odpovědí má můj blog: `https://jlibovicky.github.io//2023/02/07/Otazky-a-odpovedi-o-ChatGPT-a-jazykovych-modelech.html`

Shrnutí

1. Strojové učení – tam, kde nevíme, jak řešení naprogramovat
2. Jazykové modely – pořád hlavně předtrénování obecného modelu, který se doladí na konkrétní úlohu
3. Instruction-tuned jazykové modely – snaha vydestilovat obecné schopnosti z modelu, ale dotrénování je stále lepší

https://ufal.mff.cuni.cz/jindrich-libovicky

References I

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1409.0473.

Emily M Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Nick Couldry and Ulises A Mejias. *The costs of connection*. Stanford University Press, 2019.

Kate Crawford. *The atlas of AI*. Yale University Press, 2021.

Eva Derous and Ann Marie Ryan. When your resume is (not) turning you down: Modelling ethnic bias in resume screening. *Human Resource Management Journal*, 29(2):113–130, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

References II

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Nal Kalchbrenner and Phil Blunsom. Recurrent continuous translation models. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1700–1709, 2013.

Jindřich Libovický. Neuronové sítě a automatický překlad. *Rozhledy matematicko-fyzikální*, 94(4):30–40, 2019. ISSN 0035-9343.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *CoRR*, abs/2203.02155, 2022. doi: 10.48550/arXiv.2203.02155. URL `https://doi.org/10.48550/arXiv.2203.02155`.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtskỳ. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature communications*, 11(1):1–15, 2020.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1 (8):9, 2019.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125, 2022. doi: 10.48550/arXiv.2204.06125. URL `https://doi.org/10.48550/arXiv.2204.06125`.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01042. URL `https://doi.org/10.1109/CVPR52688.2022.01042`.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? *arXiv preprint arXiv:2304.15004*, 2023.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.