Hluboké sítě, transformery, velké jazykové modely – jemný úvod

Václav Hlaváč
České vysoké učení technické v Praze
Český institut informatiky, robotiky a kybernetiky (CIIRC)
160 00 Praha 6, Jugoslávských partyzánů 1580/3
vaclav.hlavac@cvut.cz, http://people.ciirc.cvut.cz/hlavac/

Co je umělá inteligence (AI)?

- Umělá inteligence (AI) je technologie, která umožňuje řešit počítači úlohy, které normálně vyžadují lidskou inteligenci, např. učení, porozumění jazyku, rozpoznávání obrázků, řešení úloh a rozhodování
- Vzpomeňte na slavný test, Alan Turing Test, 1950

- Edsger W. Dijkstra (1930-2002): *Otázka, zda může počítat myslet není zajímavější, než otázka, zda ponorka umí plavat*
- Umělá inteligence je částí počítačové vědy zabývající se vývojem počítačových systémů vykazujících chování podobné lidské inteligenci.

Jak se AI vyvíjela? Předběžný přehled

- Hlavní otázka:
  Jak jsme se posunuli od ručně navržených rozhodovacích pravidel v rozpoznávání k systémům jako ChatGPT?
- Jak se AI postupně vyvíjela?
  AI → rozpoznávání → strojové učení (ML) → hluboké učení → naučené reprezentace → generativní AI
- Jak pracují moderní neuronové sítě?
  vrstvy neuronů → učení → Convolutional Neural Networks (CNNs) → přenosové učení (transfer learning) → transformer
- Proč jsou vyústěním velké jazykové modely (LLM)?
  tokens (základní textové jednotky) → pozornost → transformerový blok → předpověď příštího tokenu → ChatGPT/základní (foundation) modely

Statistické modely, klasifikátory, modely v neuronových sítích

- V této přednášce se bude často využívat pojem model. Tento pojem se hojně používá v různých souvislostech, například model, domu, digitální model výrobní linky (také digitální dvojče).
- Pojem statistický model klade důraz na pravděpodobnost, rozdělení, nejistotu či odhad založený na datech. Zdůrazňuje, že daný systém reprezentuje nebo odhaduje pravděpodobnostní vztahy v datech. Tento popis sice odpovídá mnoha systémům strojového učení, zejména jazykovým modelům, avšak ne každé použití pojmu „model“ vyžaduje takovou specifikaci.
- V současném strojovém učení je pojem „model“ v rámci neuronové sítě širší a přirozenější. Označení modelu jako prvku, který předpovídá „další token“, je zcela běžné.

Rozpoznávání vs. strojové učení

Rozpoznávání (statistické)

- Nalézá pravidelnosti nebo kategorie v datech;
- Zjišťuje empiricky, co systém dělá.

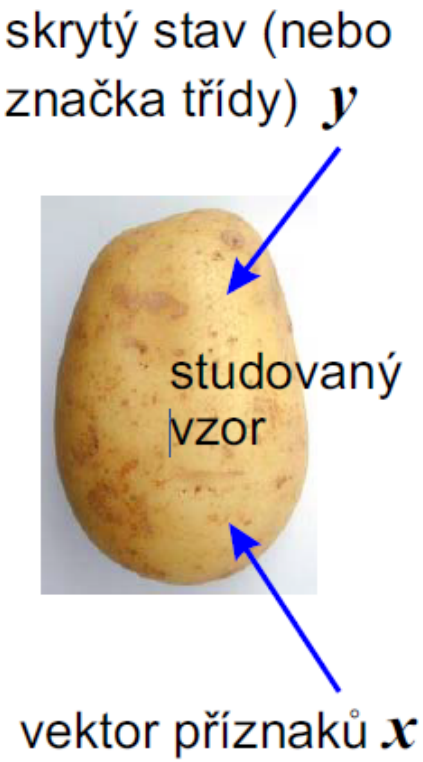
Strojové učení

- vytváří širší soubor metod, které dovolují počítači učit se z dat;
- jeden z hlavních postupů, jak se to dá udělat.

- V minulosti bylo možné systémy pro rozpoznávání vzorů vytvářet i pomocí ručně navržených pravidel a příznaků, a to bez výrazného využití učení. Moderní rozpoznávání vzorů se však obvykle realizuje pomocí strojového učení – zejména neuronových sítí.
- Strojové učení zahrnuje i řadu úkolů nad rámec rozpoznávání, jako je například předpovídání cen domů, doporučování filmů, prognózování poptávky či řízení robotů.

Rozpoznávání, základní pojmy, ilustrace “brambora”

- Studovaný vzor se analyzuje (například brambora, viz obrázek).
- Vektor příznaků $\mathbf{x} \in X$ je vektor tvořený jednotlivými pozorováními (měřeními). Vektor $\mathbf{x}$ odpovídá jednomu bodu v prostoru příznaků X.
- Skrytý stav (ve obvyklém případě přímo značka třídy) $y \in Y$ není přímo pozorovatelný. Vzory se stejnými skrytými stavy vytvářejí jednu třídu.
- Úkolem je navrhnout klasifikátor (rozhodovací pravidlo) $q: X \rightarrow Y$, které přiřazuje pozorované instance vzoru ke skrytému stavu.



$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \\ x_n \end{bmatrix}$$

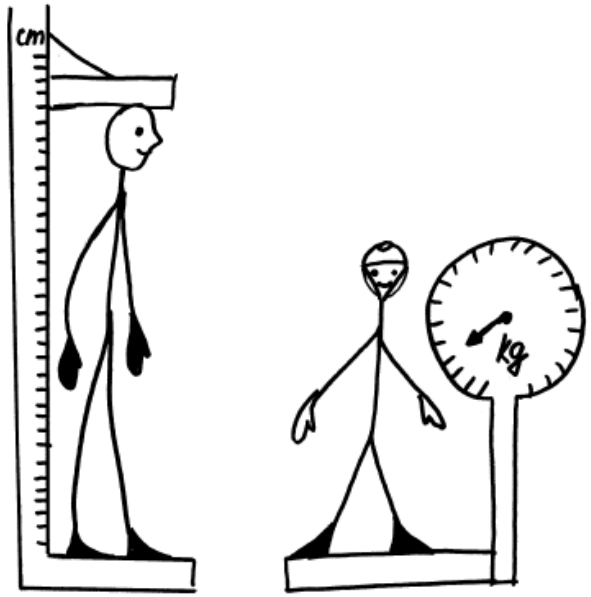
Rozpoznávání, motivační příklad, klasifikace

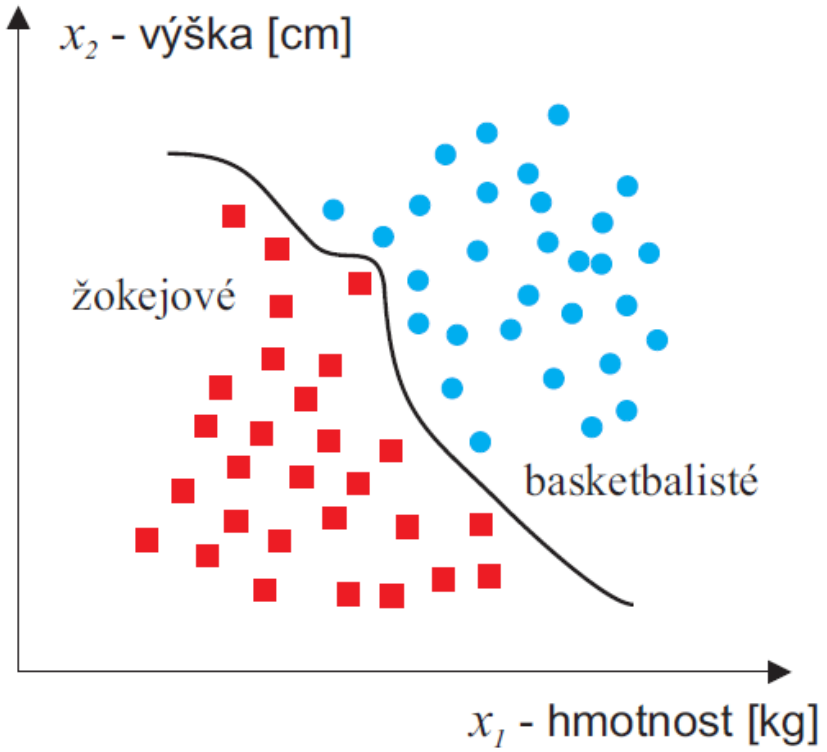
Objekt (situace) se popisuje dvěma parametry:

x -- pozorovatelný *příznak* (též pozorování).

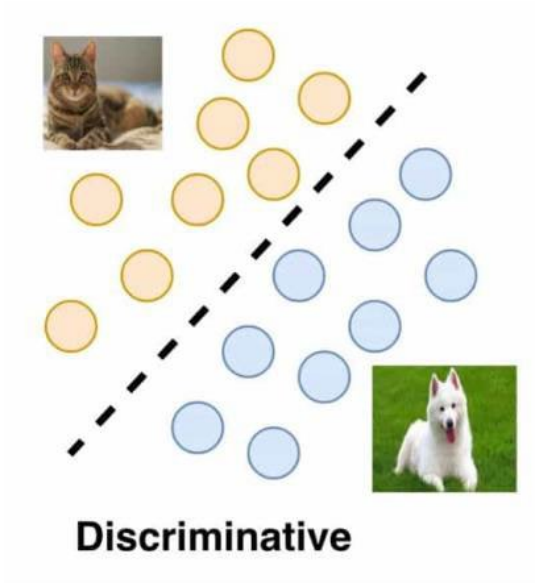
y -- *skrytý parametr* (stav, speciální případ—klasifikační třída).

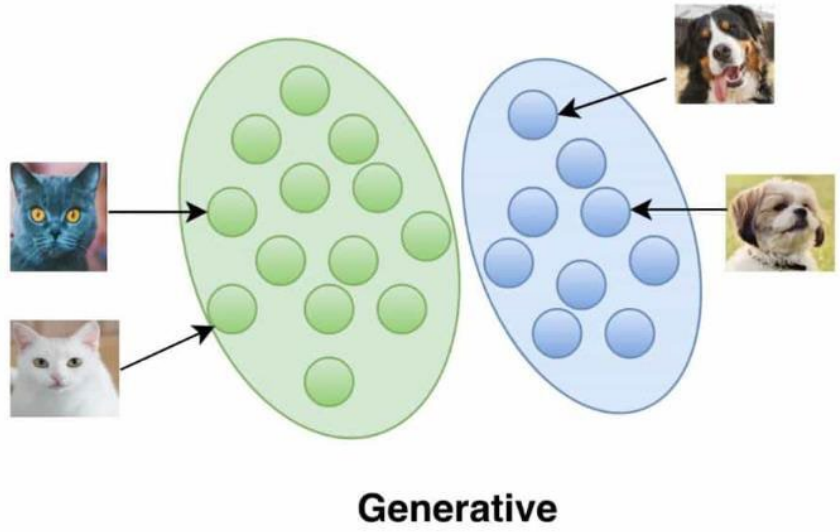
Příklad statistické rozpoznávání: žokejové a basketbalisté.





Diskriminativní vs. generativní přístup k ML



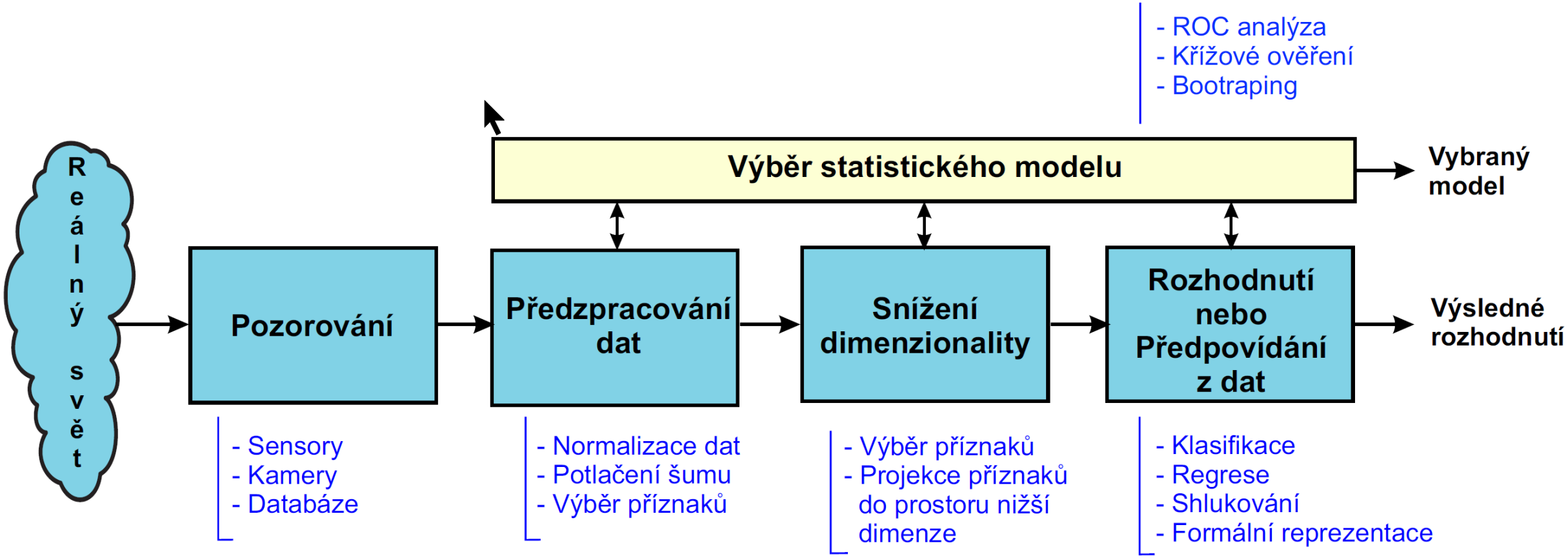


Discriminativní AI: se učí předvídat výstup ze vstupu.
Příklad: Je tento e-mail spam?

Generativní AI: se naučí dostatek ze struktury dat, aby byl schopen generovat nové správné příklady.
Example: Vytvoření odpovědi na podnět ChatBot (např. ChatCPT)

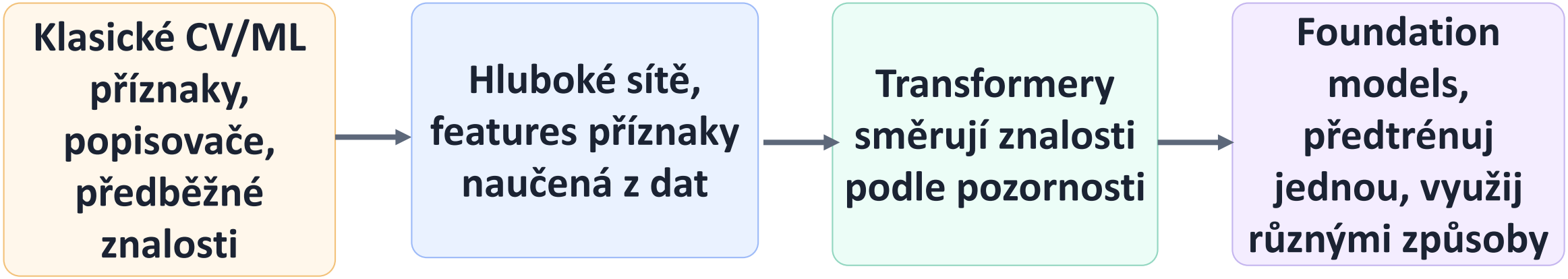
Picture: Kyu-HwanJung, Samsung

Detailnější pohled na rozpoznávání, složky



- Input: Data, trénovací (multi)-množina.
- Statitické modely a jejich parametry se empiricky učí z trénovacích dat.
- Výstupy: různorodá rozhodnutí, viz obrázek.

Myšlenkový postup: nauč se reprezentace a potom je užij



Nejde o to, že by příznaky zmizely. Jde o to, že příznaky, invariance a abstrakce úkolů se stále častěji osvojují v rámci rozsáhlých diferencovatelných systémů.

ML, hluboké učení, transformer, velké jazykové modely (LLM)

- **Strojové učení** je rozsáhlá oblast, v níž se algoritmy učí vzory z dat, místo aby byly pro každý případ výslovně naprogramovány.
- Strojové učení využívá **hluboké neuronové sítě** s mnoha vrstvami. Hluboká síť je neuronová síť s dostatečným počtem vrstev na to, aby se dokázala učit stále abstraktnější reprezentace – například hrany → tvary → objekty na obrázcích nebo znaky → slova → významy v textu.
- **Transformer** je specifický typ architektury hlubokých neuronových sítí. Jeho klíčovým principem je mechanismus pozornosti: model dokáže určit, které části vstupu jsou ve vztahu k ostatním nejvýznamnější. Transformery jsou obzvláště účinné při zpracování jazyka, ale využívají se i v oblastech, jako je zpracování obrazu, zvuku, biologie a dalších.
- Velký jazykový model (LLM) je obvykle velmi rozsáhlý model typu Transformer, který byl natrénován na obrovském množství textu za účelem předvídání a generování jazyka. Do této kategorie patří například modely řady GPT. Díky tomuto tréninku získávají schopnosti, jako je shrnování textu, překlad, zodpovídání otázek, psaní kódu či chování připomínající uvažování.

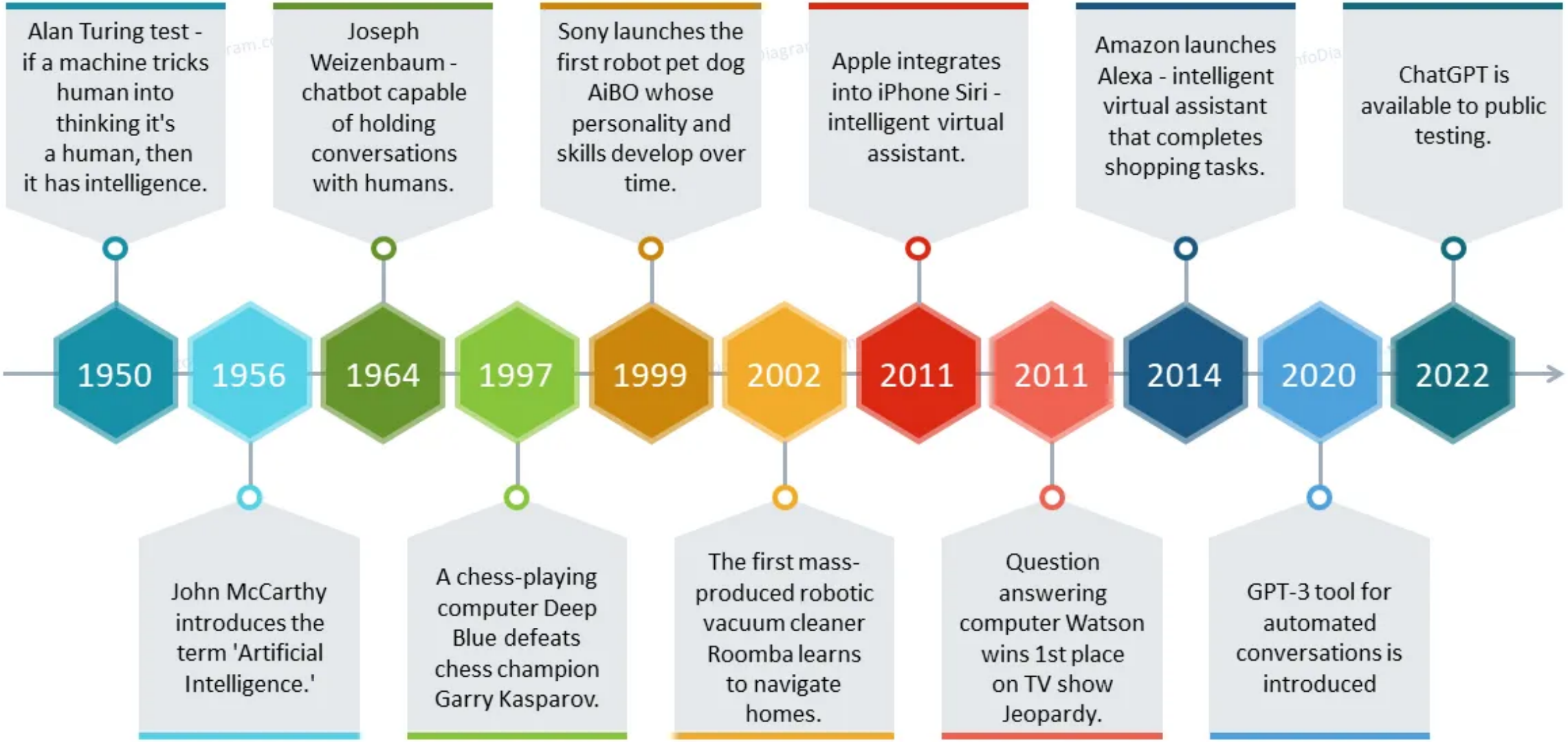
Hierarchie mezi pojmy, které jsme zavedli

- Hierachie je přibližně:
  Umělá inteligence → Strojové učení → Hluboké učení → Transformery → Velké jazykové modely
  Vnímejte to jako přibližnou hierarchii zahrnování, nikoli jako historickou posloupnost.
- Důležitý detail:
  - ne každá hluboká síť je transformer,
  - ne každý transformer je LLM.
  - LLM zvláštní použití hlubokého učení velkého rozsahu, který se obvykle buduje pomocí transformerové architektury.

Kvalitatiní posun v jazykových modelech

- Noam Chomsky (Syntactic Structures, 1957) – generativní gramatiky: formální systém pravidel pro generování gramaticky správných vět přirozeného jazyka.
- 1980.–1990. léta – nárůst používání statistických jazykových modelů založených na bigramech, trigramech; statistické přístupy později přinesly průlom ve strojovém překladu.
- 2013 Tomáš Mikolov a jeho kolegové zavedli word2vec dovolující číselnými vektory reprezentovat slova; Mikolov také přispěl k neuronovým jazykovým modelům.
- Modernní jazykové models, such as GPT (Generative Pre-trained Transformer), use a large context window to take preceding text into account when generating a response.

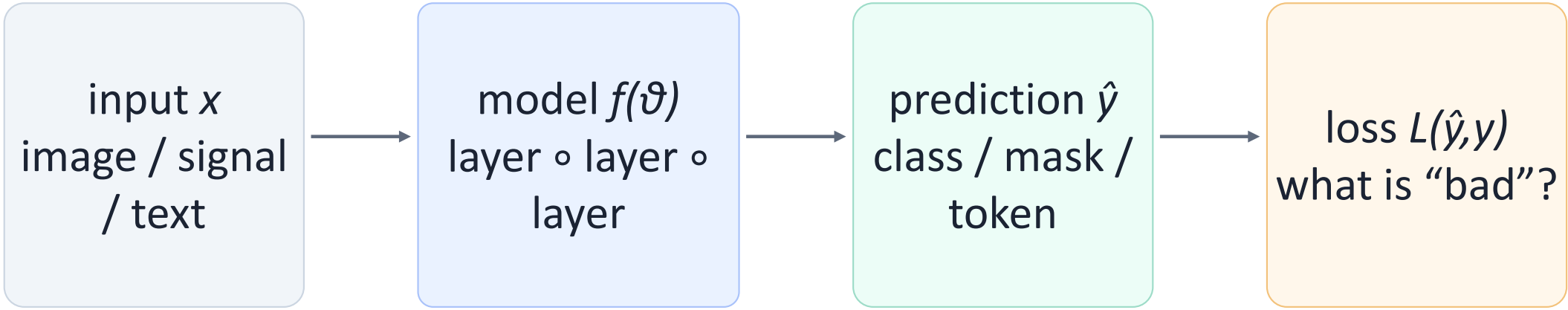
Milníky ve vývoji AI



Source: https://infodiagram.com/slides/

Návod pro hluboké učení

- Téměř veškeré učení s učitelem v oblasti hlubokého učení lze shrnout do čtyř kroků: vstup → model → predikce → měření chyby → úprava modelu.



- Zpětné navracení poskytuje $\partial L/\partial \vartheta$; optimalizátor postupně mírně upravuje parametry.
- Většina architektonických debat se týká toho, na co by měl být výpočet $f(\vartheta)$ zaměřen: lokalitu, opakování, pozornost, paměť, měřítko.

Proč se hluboké učení prudce rozjelo?

Data

veliká označkovaná data, text rozsahu celého webu, archív nemocnice

Compute

GPUs/TPUs usnadňují maticové výpočty

Triky

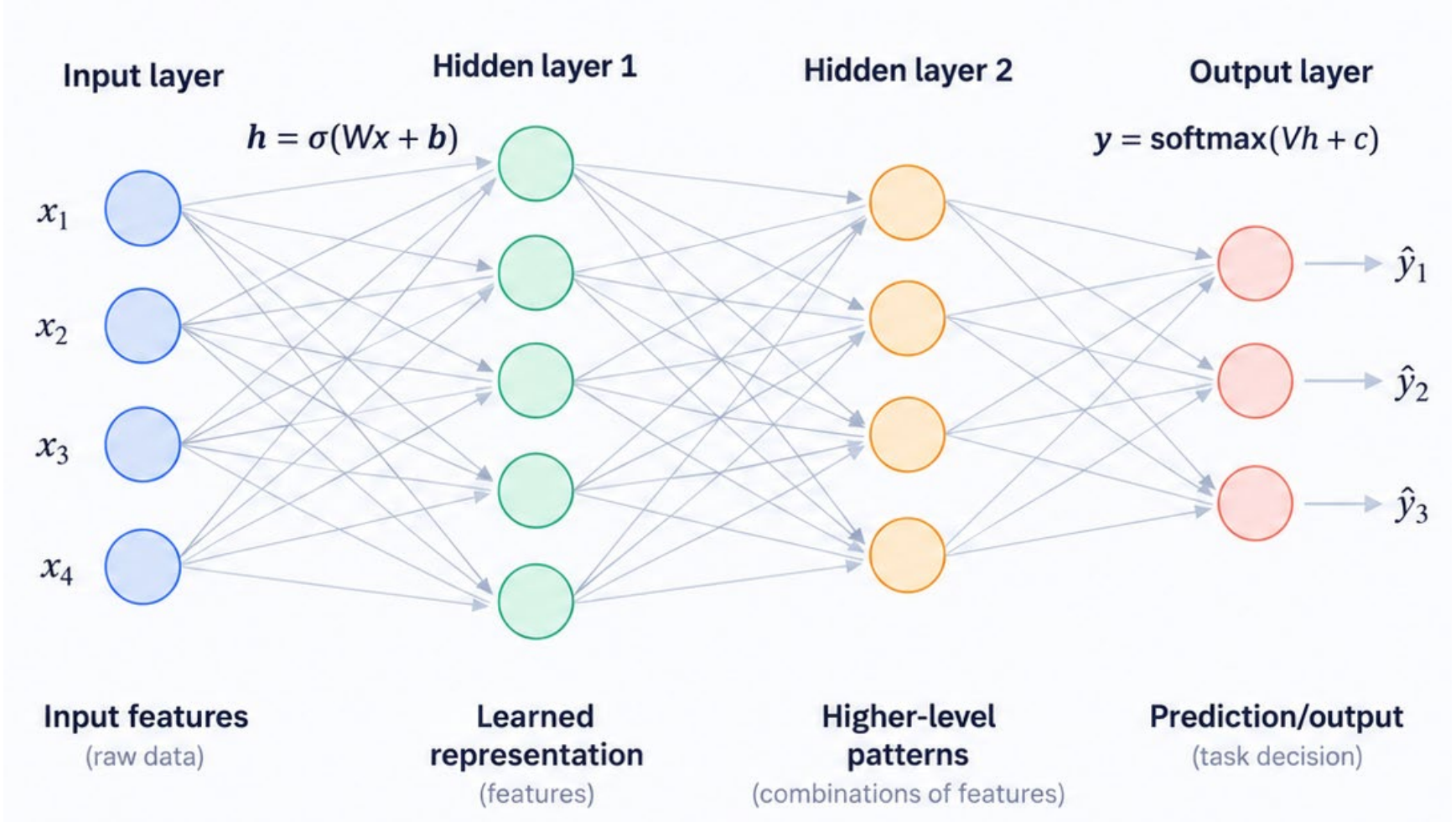
ReLU (Rectified Normalized Unit, $\mathrm{ReLU}(x) = \max(0, x),$.
normalizace, reziduální propojení ze vstupu na výstup,
dobré optimalizační nástroje

Otevřené nástroje

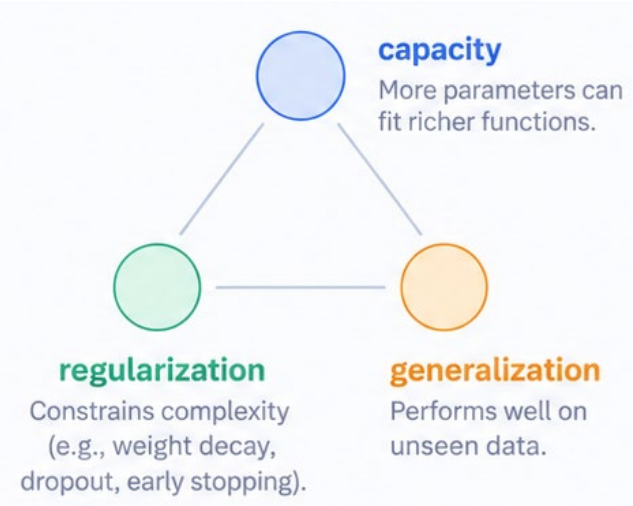
PyTorch, TensorFlow, model hubs, reproducible baselines

Koncepční podstata je stará: získejme optimalizací rodinu parametrizovaných modelů. Praktický rozdíl je v měřítku a naučených reprezentacích.

Vrstvy NN: naučené transformery + nonlinearity



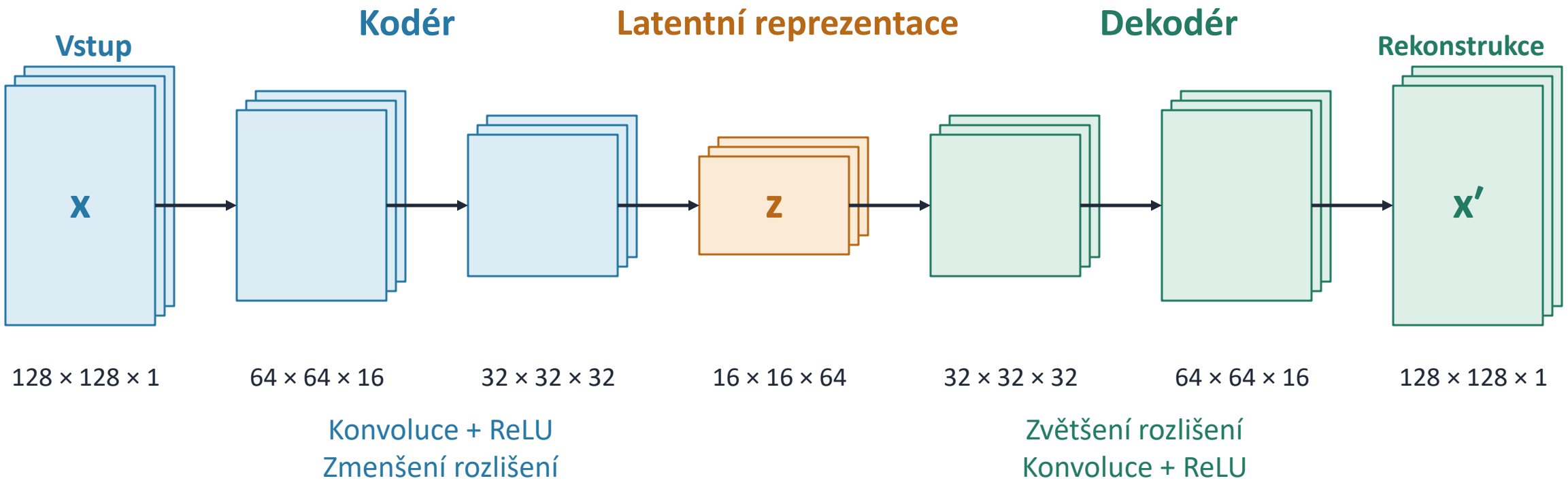
- Vrstvení vrstev vytváří reprezentace.
- Z neorganizovaných vstupů vznikají příznaky, z nich se tvoří vzory. Ty umožňují vytvořit vrstvu, která rozhoduje.



Kapacita (počet vrstev, parametrů) and regularizace ovlivňují schopnost modelu generalizovat.

CNN s kodérem a dekodérem

Autoenkodér se učí rekonstruovat svůj vstup



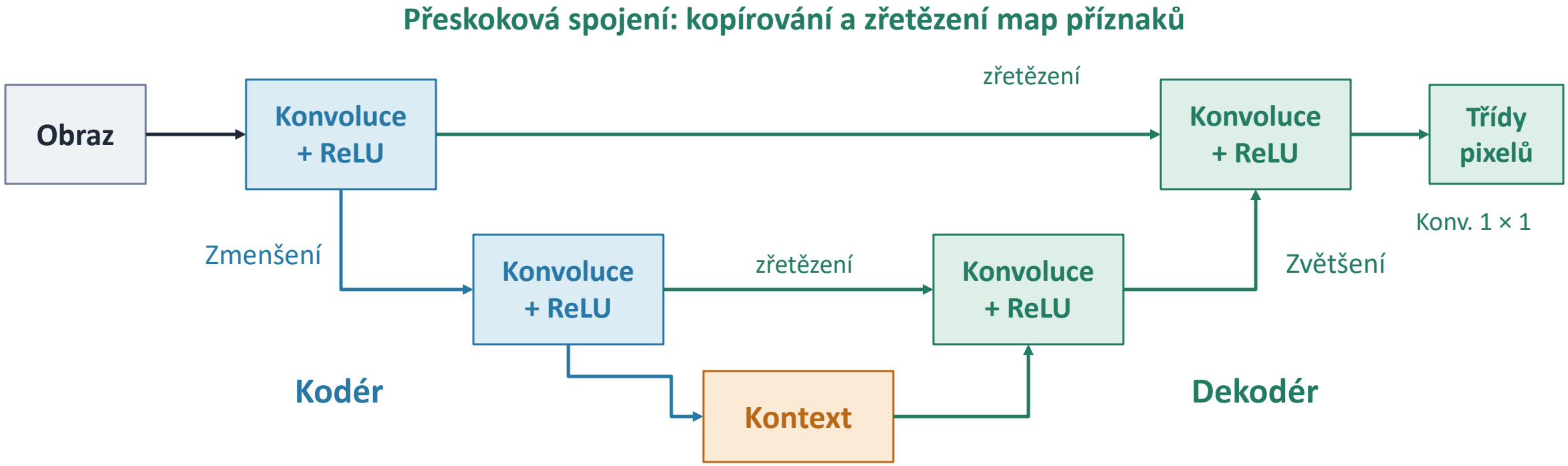
Kodér vytváří příznaky ze stále větších oblastí obrazu.

Dekodér využívá naučenou reprezentaci k rekonstrukci obrazu.

Historie: LeNet-5 (1998) klasifikoval číslice. U-Net (2015) určuje třídy pixelů.

U-Net: segmentace obrazu

CNN s kodérem a dekodérem určuje třídu každého pixelu



Kodér zachycuje kontext. Dekodér obnovuje prostorové rozlišení.

Přeskoková spojení přenášejí jemné prostorové detaily do dekodéru.

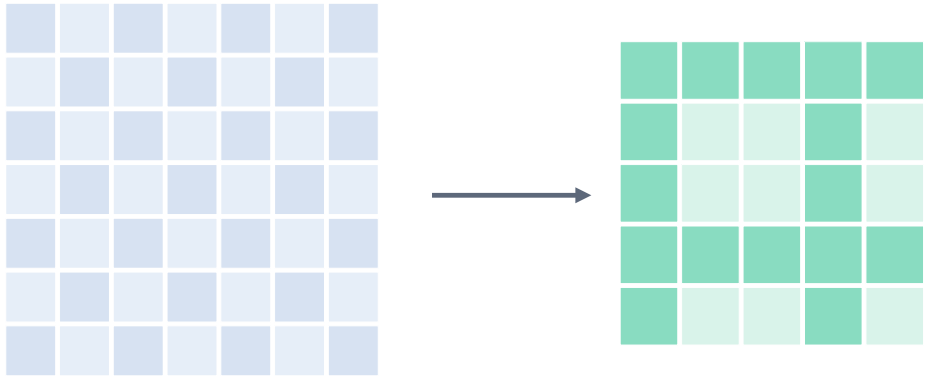
Příklad: segmentace buněk a pozadí. Zjednodušený U-Net se dvěma zmenšeními rozlišení.

CNNs for obrázky a 3D data (objemy)

lokální receptivní pole

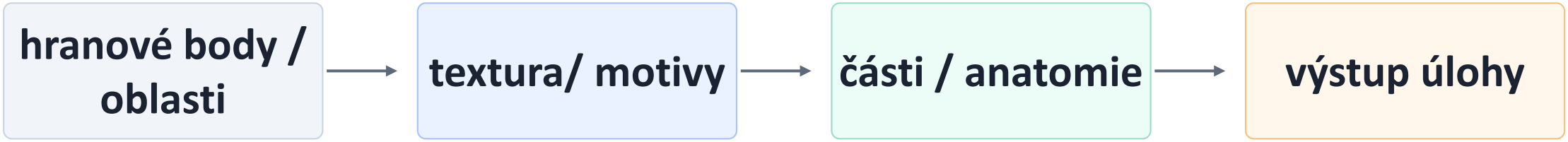
stejná konvoluční jádra všude

mapy příznaků zachovávají geometri



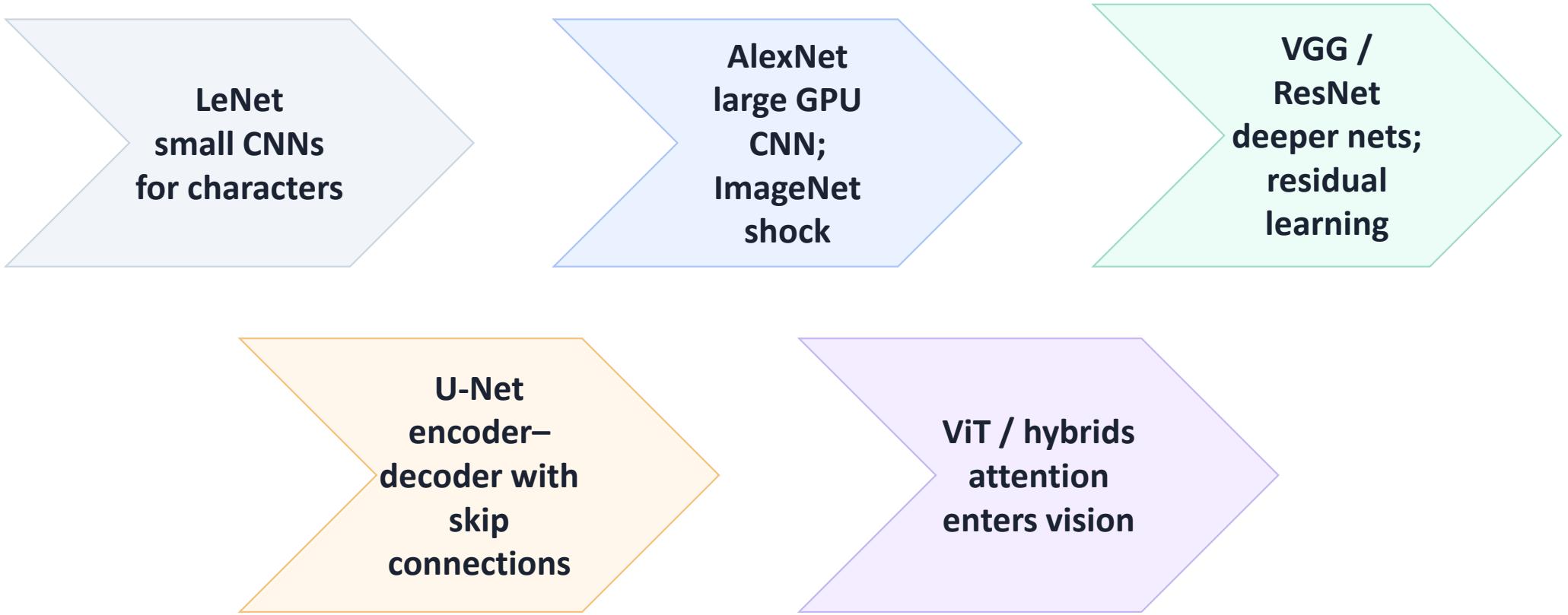
- CNNs: naučené banky filtrů se sdílenými vahami
- Konvoluce zachycuje lokální chování a je invariantní vůči posunutí. Nejsou to kouzla. Řešení operací sdružování (pooling) a posunu okna (striding) s ohledem na invariantnost a kontext.

CNN se učí hierarchii užitečných statistik



- Model se může naučit spíše jednodušší rušivou proměnnou než jev, který nás zajímá.
- Příklady: typ kamery, scanner, pozadí, styl značkování

Milníky v architekturách CNN: Co se změnilo?



Výběr ztrátové funkce je volbou při hledání modelu

Cross entropy
probabilistic classification
next-token prediction

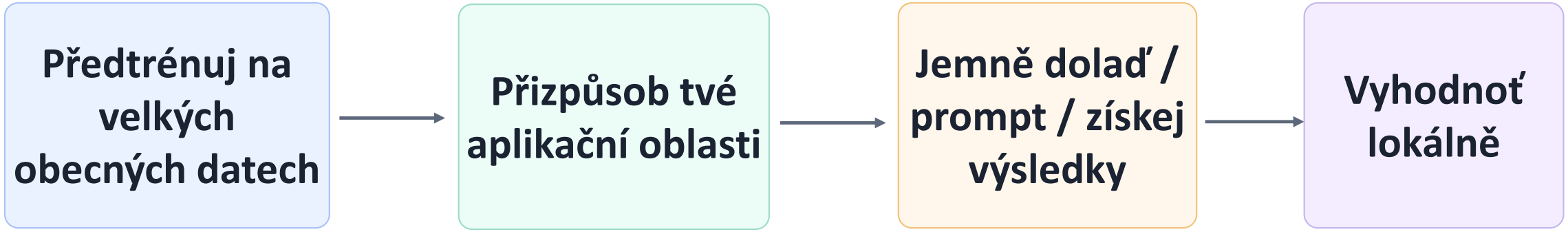
Dice / IoU
class imbalance in segmentation

Contrastive
bring matched representations together

Reconstruction
autoencoders, denoising, diffusion

- Ztrátová funkce modelu sděluje, co se považuje za úspěch.
- Metrika říká lidem, zda je tento úspěch užitečný.

Přenosové učení (transfer learning) změnilo předchozí postup



Příklad: U malých biomedicínských datových sad už otázka často nezní „mohu model natrénovat od nuly?“, nýbrž „jakou již existující reprezentaci bych měl převzít a jaké vychýlení to s sebou přinese?“

Transformery

CNN
stejné lokální operace všude

Transformer
každá textová jednotka (token)
se ptá: “Na co se mám zaměřit?”

- Od pevných filtrů k interakcím závislým na obsahu.
- Mechanismus pozornosti umožňuje každému tokenu kombinovat informace z ostatních tokenů pomocí vah závislých na obsahu.
- Je to mocný nástroj pro jazyk, ale uplatní se i u obrázků, proteinů, časových os elektronických zdravotních záznamů a multimodálních dat.

Sebepozornost v jednom řádku

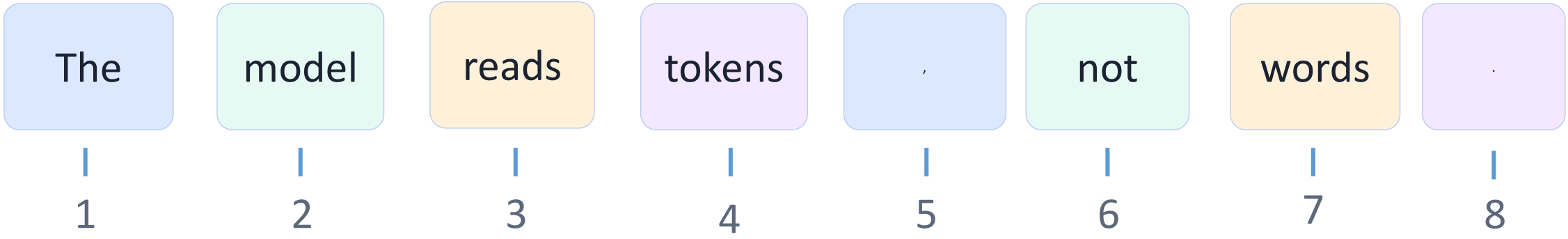
Otázka
Co hledám?

Klíč
Co nabízím?

Hodnota
Jakou znalost posílám?

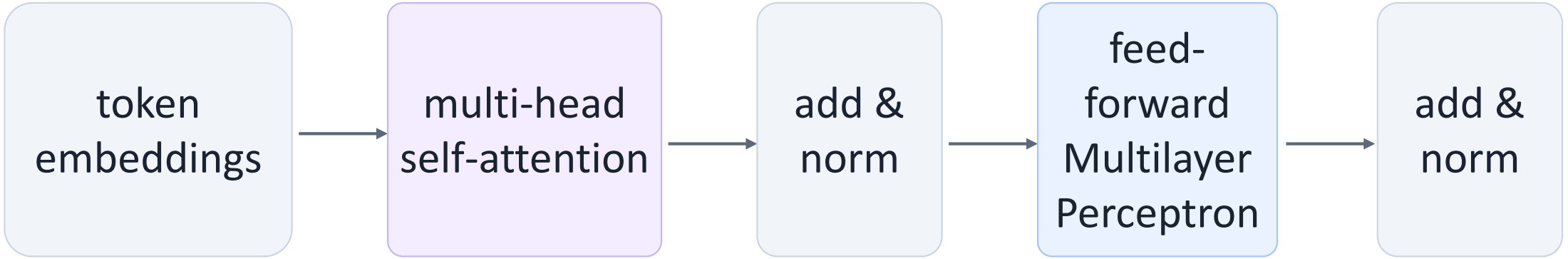
- $Attention(A, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$
  - Q = Query (otázka)— what a token is looking for
  - K = Key (klíč) — what a token offers for matching
  - V = Value (hodnota) — the information carried by that token
  - d_k = dimenze vektorů klíče a/nebo otázky;
- Mechanismus pozornosti s více hlavami (multi-head attention) tento postup opakuje pro několik naučených podprostorů, takže různé hlavy mohou modelovat různé vztahy.

Transformery a základní jazykové jednotky (tokeny)



- Transformery pracují s základními jazykovými jednotkami (tokeny) a nikoliv s hrubými významy
- Obrázky se mohou stát tokeny: oblasti ve Vision Transformer, oblasti v detektory nebo vnoření z CNN. Vnoření (embedding) je vektor čísel (příznaky), který CNN vytáhla z obrázku.

Blok transformerů



- Blok transformerů je: mechanismus pozornosti plus lokální výpočty
- Reziduální propojení (residual paths) a normalizace umožňují trénování hlubokých zásobníků (deep stacks, mnohovrstvených sítí).

Tři „osobnostní rysy“ transformerů

Encoder-only
BERT style
understand/label
inputs

Decoder-only
GPT style
generate next token

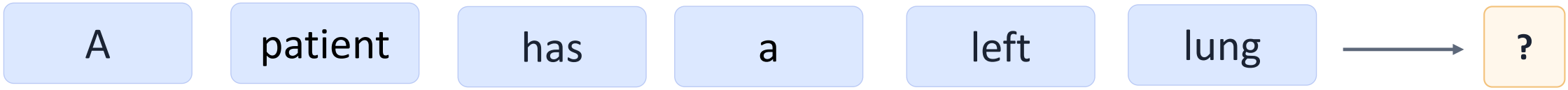
Encoder–decoder
T5/translation style
input → output
sequence

Poučení: Vyber si skupinu transformerů podle struktury řešené úlohy a nikoliv podle módnosti.

LLMs a základní (foundation) modely

- Co se změní, když se modely typu Transformer trénují na celém internetu?
- Postup trénování, nikoliv jen model
  1. Předtrénování: nauč se z masivní sady dat — např. předpověď příštího tokenu pro velké jazykové modely pro předvídání ve stylu GPT (Generative Pre-trained Fransformer).
  2. Ladění instrukcí: nauč se sledovat „task-like prompts“
  3. Ladění preferencí / bezpečnosti: přispívá k užitečnosti a zakazuje některá chování
  4. Použití nástroje / vybavování: propoj s vnější znalostí a akcemi
- Model se učí statistické struktuře; úspěch nástroje také závisí na podnětech (promts), okně poskytujícím kontext, nástrojích a policy (tj. pravidlech a doporučeních řídicích, jak se AI systém chová).

Předvídání příští základní jazykové jednotky (tokenu)



- Pravděpodobnost(příští token | předchozí tokeny)
- Tím, že se model učí pokračovat v textu, si osvojuje gramatickou strukturu, sémantické vztahy, styly, vzorce úkolů, četné faktické souvislosti i určité vzorce připomínající uvažování. Avšak předpovídání následujícího tokenu nezaručuje jeho pravdivost.
- Pro (biomedicínské) účely přistupujte k velkým jazykovým modelům (LLM) jako k asistentům, nikoli jako k autoritám.

ChatGPT

- Interaktivní systém pro konverzaci a generování textu.
- Využívá pravděpodobnostní jazykový model k předpovědi následujícího tokenu na základě kontextu.
- Udržuje kontext v průběhu konverzace.
- Dokáže se řídit pokyny v přirozeném jazyce.

- Může chybovat a vymýšlet si znalost
- ChatGPT byl spuštěn společností OpenAI v listopadu 2022.
- Chatbot je víc než jen základní jazykový model: na jeho odpovědi mohou mít vliv také systémové pokyny, nástroje, vyhledávání a další součásti.

LLM je pouze jednou součástí systému umělé inteligence.

Analogie: LLM poskytuje něco jako „kognitivní jádro“, zatímco nástroje a agenti dávají systému schopnost vnímat a jednat.

- LLM — kognitivní jádro: zpracovává jazyk a kontext a navrhuje, co říci nebo udělat.
- System prompt — role a pravidla: určují modelu, jak se má chovat.
- Context — pracovní paměť: znalosti dostupné během probíhající interakce.
- RAG (Retrieval-Augmented Generation) — přístup k referenčním materiálům: přináší do kontextu relevantní externí znalost.
- Tools and agents — schopnost jednat: vyhledávat, provádět výpočty, číst soubory, odesílat zprávy, volat software.
- MCP (Model Context Protocol) — standardní připojení k nástrojům a datům.

RAG: poskytne modelu relevantní znalost

- RAG — Retrieval-Augmented Generation
- Otázka → Vyhledávání / Získávání → Relevantní dokumenty → LLM → Odpověď
- Interní znalosti LLM mohou být neúplné, zastaralé nebo nedostatečně specializované.
- RAG získává relevantní externí informace a dává je do kontextu modelu.
- Model poté vygeneruje odpověď s využitím otázky i získaného materiálu.
- RAG zlepšuje ukotvení v datech, ale nezaručuje správnost.
- Příklad z biomedicíny: „Co říká protokol naší nemocnice o tomto zákroku?“ → vyhledání příslušného protokolu → jeho předání modelu LLM → vygenerování odpovědi.

Agenti: od odpovídání k činnosti

- Agent může využít LLM k výběru a provádění akcí.
- Cíl uživatele → LLM → výběr nástroje → pozorování výsledku → rozhodnutí o dalším kroku → … → výsledek
- Potom
  - Chatbot hlavně odpovídá.
  - Agent dokáže plánovat, používat nástroje, sledovat výsledky a pokračovat, dokud není úkol dokončen.
  - Typické nástroje: vyhledávání na webu, databáze, spouštění kódu, e-mail, kalendář, laboratorní či podnikový software.
  - Větší míra autonomie s sebou přináší i vyšší potřebu oprávnění, ověřování a lidského dohledu.

MCP — Model Context Protocol

- Standardizovaný způsob, jakým se aplikace využívající umělou inteligenci mohou připojit k externím nástrojům a datům.
- Aplikace využívající AI ↔ MCP ↔ soubory / databáze / webové služby / nástroje

System Prompt (≈ instrukce)

- Pokyny poskytnuté aplikací, které definují roli, chování, priority a omezení modelu.
- System prompt → User prompt → Retrieved/context information → Model response

Vynikající přednášky a kursy k naučení

- CS231n, Stanford: CNNs for visual recognition; especially lectures on CNNs, architectures, detection/segmentation
- MIT 6.S191 / MIT 6.7960: Modern deep learning, transformers, language models, scaling, with slides and videos.
- Stanford CS25 Transformers United: Strong transformer overview and guest lectures, including Karpathy's intro.
- Karpathy: Let's build GPT: Hands-on from-scratch GPT; best after basic PyTorch/tensor comfort.
- Hugging Face LLM Course: Practical transformer/LLM course with runnable examples.
- Full Stack LLM Bootcamp: Applied LLM systems: prompting, RAG, evaluation, deployment.

Závěrečná stručná rekapitulace

- Deep learning
  učí se reprezentace z dat, namísto aby spoléhal především na ručně navržené příznaky.
- CNNs:
  zahrnutí principu lokality a sdíleného výpočtu – obzvláště užitečné pro obrázky.
- Transformers:
  Využijte pozornost k tomu, aby prvky komunikovaly v souladu s obsahem.
- LLMs:
  Jedná se o rozsáhlé jazykové modely založené na architektuře Transformer, které byly trénovány především pomocí predikce.
- Hlavní poučení:
  Moderní umělá inteligence kombinuje data, naučené reprezentace a škálovatelný výpočetní výkon; působivé chování však neznamená neomylnost ani lidský způsob porozumění.
- Modern AI systems = modely + kontext + retrieval + nástroje + interakce.