Obsah

1. Definujte strojové učení a vysvětlete rozdíl mezi strojovým učením a klasickým programováním.

- Oblast umělé inteligence, která se zabývá vývojem algoritmů a technik, které umožňují počítačům se učit z dat a zkušeností

Klasické programování:

- Programátor ručně píše instrukce, které určují, co se má vykonat
- Programy jsou navrženy tak, aby splňovali podmínky a prováděly specifické úkoly na základě přesně daných pravidel

Strojové učení:

- Počítač sám sebe učí na základě dat
- Počítači jsou předložena data, ze kterých se učí
- Tento proces umožňuje počítačům adaptovat se na nové situace

2. Popište strojové učení s učitelem a bez učitele, uveďte příklady metod.

S učitelem:

- Vstupním datům jsou přiřazeny správné výstupy – labeled data
- Existuje okamžitá zpětná vazba
- Cílem je např. najít klasifikační nebo regresní model pro predikci výstupů

Bez učitele:

- Vstupním datům nejsou přiřazené výstupy
- Zpětná vazba chybí
- Cílem je např. odhalit v datech skrytou strukturu

3. Vysvětlete princip jedné metody strojového učení s učitelem a uveďte konkrétní úlohu, v níž lze metodu využít.

Princip logistické regrese:

- Klasifikační algoritmus
- K předpovídání pravděpodobností, že určitý vstup patří do jedné z kategorií
- Na základě vstupních dat se model naučí vztah mezi těmito daty a výstupní třídou
- Model se trénuje na sadě, kde jsou známé vstupy i výstupy
- Používá logistickou funkci sigmoid, která převádí pravděpodobnost na rozmezí od 0 do 1

Úloha:

- Detekce spamu v mailu
    - Vstupní data mohou být klíčová slova/výrazy obsažená ve spam mailech
    - Logistická regrese se na základě těchto slov naučí rozpoznávat vzory, které jsou typické pro spam

4. Vysvětlete princip jedné metody strojového učení bez učitele a uveďte konkrétní úlohu, v níž lze metodu využít.

Princip k-means:

- Používá se k rozdělení dat do k počtu skupin tak, aby data ve skupině byla co nejpodobnější
- Na začátku se vyberou náhodně centroidy a algoritmus potom:
    - Přiřadí – datový bod se přiřadí ke skupině s nejbližším centroidem
    - Aktualizace – vypočítá se nový centroid každé skupiny jako průměr všech bodů v skupině
- Tento proces se opakuje, dokud se skupiny nadále nemění, nebo nedosáhneme maximálního počtu iterací

Úloha:

- Segmentace zákazníků v marketingu
    - Data mohou zahrnovat informace o chování zákazníků – četnost nákupů, útrata, kategorie produktů…
    - Cílem je rozdělit zákazníky do skupin s podobnými nákupovými vzory, i když předem nevíme, kolik skupin existuje ani jak vypadají

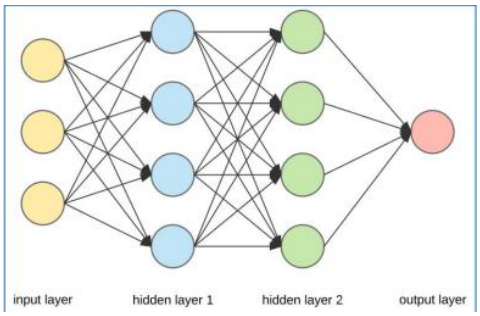
5. Popište rozdíly mezi biologickou a umělou neuronovou sítí.

- Biologická – mozek
    - Miliardy neuronů spojené synapsemi
    - Neurony komunikují pomocí elektrochemických signálů
    - Každý neuron má velmi složitou strukturu a reaguje na chemické látky
    - Přenáší informace pomocí elektrických impulsů
    - Učí se prostřednictvím změn synaptické síly, často ne zcela pochopeným způsobem
    - Velmi efektivní z hlediska spotřeby energie
    - Extrémně flexibilní a dokáže se rychle přizpůsobit situacím
- Umělá
    - Z jednoduchých matematických modelů neuronů – pracuje s číselnými hodnotami
    - Spoje mezi neurony – váhy, jsou vyjádřeny čísly, která určují sílu signálu
    - Struktura je výrazně zjednodušená
    - Přenos signálu je matematický – využívá se součet vážených vstupů a aktivační funkce
    - Učí se pomocí algoritmů
    - Vyžaduje výrazně vyšší výpočetní výkon a energii
    - Potřebuje velké množství dat a času na trénování – méně odolné vůči neznámým vstupům

6. Popište základní architekturu a fungování umělé neuronové sítě.

Architektura:

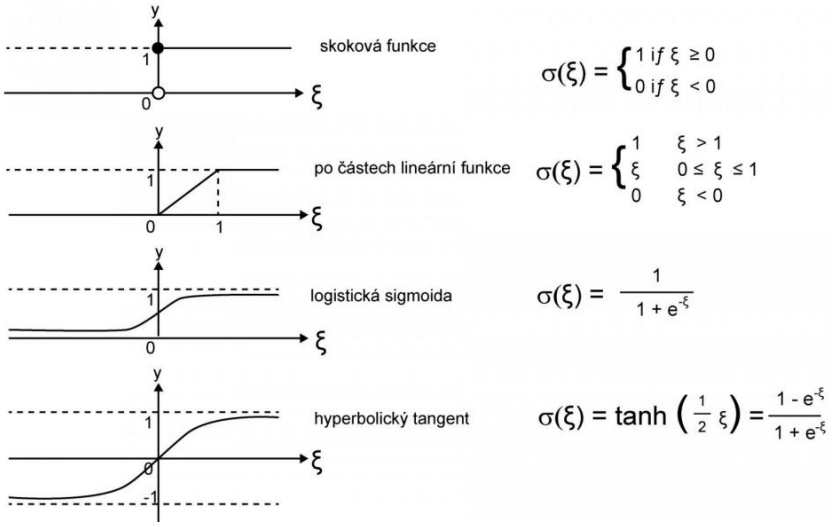
- Vstupní vrstva
    - Přijímá vstupní data
    - Každý neuron odpovídá jednomu prvku vstupu
- Skryté vrstvy
    - Dochází k výpočtům a transformaci dat pomocí vah a aktivační funkce
    - Čím více vrstev, tím „hlubší“ je síť
- Výstupní vrstva
    - Produkuje finální výstup sítě
    - Počet neuronů závisí na typu úlohy



Fungování:

- Propagace vstupu
    - Každý neuron předá svou hodnotu do skryté vrstvy
    - Ve skryté vrstvě spočítá vážený součet vstupů
    - Výsledek se transformuje pomocí aktivační funkce
    - Výstup se posílá do další vrstvy
- Výstup a rozhodnutí
    - Ve výstupní vrstvě se produkuje výsledek
- Učení
    - Síť se učí pomocí zpětného šíření chyb a optimalizačního algoritmu
    - Porovná predikovaný výstup se skutečným výstupem a upraví váhy tak, aby příště byla chyba menší
- Testování
- Používání sítě
    - Řešení úlohy naučenou sítí
    - Váhy zůstávají pevné, nebo se průběžně adaptují

7. Jmenujte a nakreslete aktivační funkce nejčastěji používané v neuronových sítích.



8. Popište perceptron a jeho učení.

- Frank Rosenblatt (1957)
- Základní model dopředné neuronové sítě, s učením, obsahuje pouze jeden neuron
- Zjištění, že je jen omezeně využitelný – jen pro lineárně separabilní úlohy
- Používá se především pro binární klasifikaci

Učení:

- Učí se z dat tak, že upravuje své váhy podle chyb, které dělá
    - Zadej vstup a spočítej výstup perceptronu
    - Porovnej výstup s očekáváním
    - Aktualizuj váhy a bias, pokud je predikce chybná
    - Opakuj, dokud se nenaučíš správně klasifikovat

9. Vysvětlete pojmy klasifikace a klasifikátor, popište typy klasifikace obrazu.

Klasifikace:

- Metoda učení s učitelem – proces kategorizace do tříd
- Obdoba predikce, kdy hodnoty závislé proměnné y nabývají malého počtu diskrétních hodnot
- Nejjednodušší je binární klasifikace – 0: negativní třída, 1: pozitivní třída

Klasifikátor:

- Algoritmus, který mapuje vstupní data na kategorie

Typy:

- Binární klasifikace

    - Zařazení do jedné ze dvou kategorií
- Klasifikace do více tříd
    - Zařazení do jedné z několika kategorií
- Multi-label klasifikace
    - Zařazení obrazu do více kategorií
- Hierarchická klasifikace
    - Zařazení do více úrovní hierarchie
- Jemnozrnná klasifikace
    - Odlišení podobných kategorií, obrázky ve vysokém rozlišení a složité modely
- Zero-Shot klasifikace
    - Klasifikace snímků, které model dosud nezpracoval s využitím sémantických informací o nových kategoriích, výsledkem je pochopení vztahu mezi známými kategoriemi a novou kategorií

10. Vysvětlete rozdíl mezi detekcí, lokalizací a segmentací (ve vztahu k obrazovým datům).

Detekce:

- Identifikuje a lokalizuje více objektů v rámci obrazu, přičemž pro každou detekovanou položku poskytuje jak štítky, tak prostorové polohy

Lokalizace:

- Zaměřuje se na hlavní objekt s ohraničujícím rámečkem

Segmentace:

- Přiřazuje každý pixel v obraze určité třídě nebo instanci objektu

11. Vysvětlete, jak (pomocí čeho) se hodnotí úspěšnost klasifikačního modelu do více tříd.

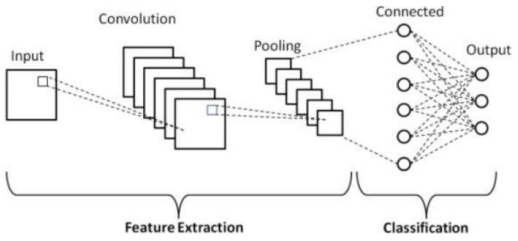
- Confusion matrix
    - Tabulka, kde řádky představují skutečné třídy a sloupce předpovězené
- Accuracy
    - Udává podíl správně klasifikovaných vzorků ze všech
- Precision
    - Kolik označených vzorků pro danou třídu patří do té třídy
- Recall
    - Kolik vzorků z dané třídy bylo rozpoznáno
- F1-skore
    - Harmonický průměr mezi precision a recall
- Cross-validation
    - Model je testován na různých podmnožinách dat

12. Vysvětlete, proč klasické algoritmy strojového učení (např. náhodné lesy nebo metoda k nejbližších sousedů) nejsou vhodné ke zpracování obrazových dat.

- Vysoká dimenzionalita obrazových dat
  - Obrázky mají mnoho vstupních znaků – pixelů
  - Klasické algoritmy neumějí s tak rozsáhlými daty efektivně pracovat
    - Jsou náchylné na přeučení
- Nevyužívají prostorovou strukturu obrazu
  - V obrazech jsou důležité lokální vzory a struktury
  - Lesy pracují s jednotlivými vstupy nezávisle a neumí zachytit vztahy mezi sousedními pixely
  - Konvoluční sítě zpracovávají obraz jako celek a umí rozpoznat vzory
- Předzpracování dat
  - Klasické algoritmy vyžadují extrakci příznaků – tedy ruční výběr a zpracování vlastností z obrazu
  - Moderní neuronové sítě se to ale mohou naučit automaticky bez zásahu
- Špatná škálovatelnost

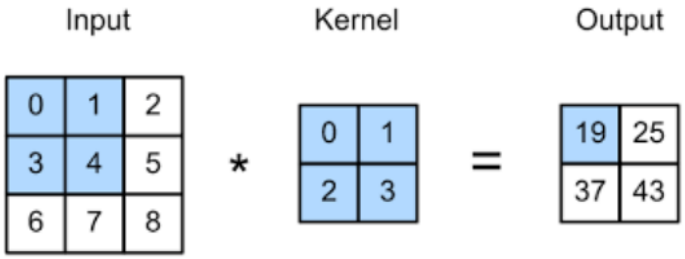
13. Vysvětlete princip konvolučních neuronových sítí.

- Druh neuronových sítí, vhodný pro klasifikaci obrazu
- Používají konvoluční vrstvy k automatickému a adaptivnímu učení prostorových hierarchií prvků ze vstupních obrázků



14. Vysvětlete pojem konvoluce (ve vztahu ke zpracování obrazu).

- Zahrnuje vynásobení hodnot jádra – specializované filtry, s původními pixelovými hodnotami obrazu a následné sečtení výsledků



$(0 \times 0) + (1 \times 1) + (3 \times 2) + (4 \times 3) = \mathbf{19}$
$(1 \times 0) + (2 \times 1) + (4 \times 2) + (5 \times 3) = \mathbf{25}$
$(3 \times 0) + (4 \times 1) + (6 \times 2) + (7 \times 3) = \mathbf{37}$
$(4 \times 0) + (5 \times 1) + (7 \times 2) + (8 \times 3) = \mathbf{43}$

15. Popište architekturu konvoluční neuronové sítě.

- Konvoluční vrstvy
  - Aplikují na vstupní obraz řadu filtrů – jader
  - Každý filtr vytváří bodový součin mezi filtrem a místními oblastmi vstupu
    - Tato operace vytvoří mapy prvků, které zachycují různé aspekty obrazu
- Sdružovací vrstvy
  - Ke zmenšení prostorových rozměrů map prvků
    - Pomáhá snižovat výpočetní složitost
    - Zabraňuje nadměrnému přizpůsobení
    - Nejběžnější max-pooling
- Počet parametrů ve vrstvě závisí na velikosti jádra a počtu filtrů
  - Každý neuron přijímá vstupy z místní oblasti předchozí vrstvy
  - Receptivní pole
    - Pohybují se nad vstupem, počítají bodové součiny a vytvářejí konvolvovanou mapu prvků jako výstup
  - Obvykle pak tato mapa prochází aktivační funkce rektifikované lineární jednotky

16. Popište architekturu sítě AlexNet.

- Jednalo se o první architekturu, která využívala GPU ke zvýšení tréninkového výkonu
- AlexNet se skládá z 5 vrstev konvoluce, 3 vrstev s maximálním sdružováním, 2 normalizovaných vrstev, 2 plně propojených vrstev a 1 vrstvy SoftMax
- Každá konvoluční vrstva se skládá z konvolučního filtru a nelineární aktivační funkce ReLU

17. Popište metody segmentace obrazu.

- Regionální metody
  - Segmentují na základě podobnosti pixelů
  - Vhodné pro obrazy s homogenními oblastmi
- Metody založené na hranicích
  - Detekují hrany v obraze pomocí gradientních operátorů
  - Pro obrazy s jasně definovanými hranicemi

- Metoda aktivních kontur
    - Používá se iterativní algoritmus, který deformuje počáteční křivku tak, aby se přizpůsobila hranicím objektu v obraze
    - Vhodné pro segmentaci složitých objektů
- Segmentace rozvodím
    - Považuje obraz za topografickou mapu, kde intenzita pixelů představuje výšku a provádí zaplavení mapy
    - Vhodné pro obrazy s různými intenzity jasu
- Metody hlubokého učení
    - Používají CNN, jsou velmi přesné
    - Vhodné pro autonomní řízení, nebo lékařské zpracování

18. Uveďte hlavní aplikační oblasti, v nichž se uplatňují metody segmentace obrazu.

- Autonomní vozidla
- Analýza lékařského zobrazení
- Analýza satelitních snímků
- Bezpečnostní systémy
- Moderování obsahu na sociálních mediích
- Chytré zemědělství
- Průmyslová kontrola

19. Vysvětlete hlavní problémy spojené s tvorbou vlastního modelu učení z obrazových dat.

- Je zapotřebí připravit velký objem předzpracovaných dat – velké časové i finanční náklady
- Trénování vlastního modelu je náročné na výkon

Řešení:

- Použití předem trénovaných modelů, které lze upravit a doladit

20. Popište databázi ImageNet a její význam pro rozvoj metod učení z obrazových dat.

- Rozsáhlá databáze anotovaných obrázků pro trénování a testování algoritmů pro rozpoznávání a klasifikaci objektů
- Obrázky jsou ručně anotované – vysoká kvalita a přesnost dat
- Klíčová databáze pro rozvoj CNN

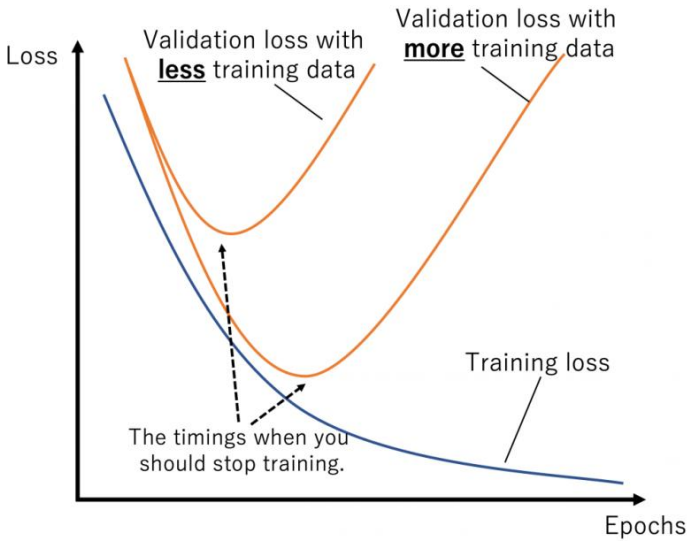
21. Vysvětlete princip přeneseného učení (transfer learning).

- Výběr před trénovaného modelu
    - Trénovaného na velkých datasetech
- Odstranění poslední vrstvy
    - Která obvykle slouží k predikci pro specifický úkol

    - Tím vznikne model, který funguje jako extraktor charakteristik
- Přidání nových vrstev
    - Čímž se model přizpůsobí pro nové zadání
- Doladění
    - Tedy trénink nových vrstev na nové datové sadě
    - Při ladění se váhy před trénovaného modelu zmrazí
        - Pouze váhy nových vrstev se aktualizují
    - Váhy modelu představují jeho klíčový aspekt
    - Doladěním modelu můžeme upravit naučené funkce tak, aby lépe odpovídaly nové sadě dat a zlepšily výkon modelu

22. Vysvětlete význam hyperparametrů při trénování neuronové sítě.

- Velikost dávky
    - Určuje počet trénovacích příkladů, které jsou při aktualizaci vah sítě zpracovány najednou
    - Menší dávky
        - Častější aktualizace může vést k rychlejší konvergenci, ale také k větší variabilitě, což může způsobit nestabilitu
    - Větší dávky
        - Aktualizace vah je stabilnější
        - Vyžaduje více paměti
        - Pomalejší konvergence
- Počet epoch
    - Určuje, kolikrát se proces učení zopakuje
    - Nastavuje se na základě sledování výsledků na validační sadě
    - Více epoch
        - Může zlepšit výkon modelu, ale roste riziko přeučení
    - Méně epoch
        - Může vést k nedotrénování modelu



23. Vysvětlete, v čem spočívají limity přeneseného učení.

- Neshoda datových sad
  - Přenesené učení funguje nejlépe, když jsou zdrojová a cílová doména podobné
  - Řešením je doladění na cílovém datasetu
- Velikost a kvalita dat
  - Nedostatek dat pro cílovou úlohu může snížit výkon modelu
  - Řešením je augmentace datové sady
- Přetrénování
  - Pokud je cílový dataset malý, model se „až moc naučí“ a ztratí schopnost generalizace
  - Řešením je sledování výkonu
- Výpočetní náročnost
  - Při použití hlubokých modelů anebo omezených výpočetních zdrojů
  - Řešením je optimalizace
- Interpretovatelnost
  - Před trénované modely mohou být obtížně interpretovatelné
    - Může být problém tam, kde je nutné chápání rozhodovacích procesů modelů
  - Řešením jsou techniky pro vizualizaci aktivací nebo metody vysvětlení rozhodnutí modelu

24. Vysvětlete pojem augmentace obrazových dat, uveďte běžné metody.

Augmentace:

- Umělé rozšíření velikosti datasetu transformací původních obrázků pomáhá zlepšit generalizaci modelu a zamezit overfittingu

Metody:

- Rotace, překlopení, oříznutí, změna měřítka, posun, změna jasu/kontrastu, přidání šumu, zkreslení

25. Popište model VGG16.

- Základní vysoce přesný model pro různé úlohy klasifikace obrazů, detekce objektů a segmentace obrazů

Architektura:

- Vstup: pevná velikost vstupního obrazu 224x224 RGB
- Konvoluční vrstvy: 13 vrstev s malými 3x3 filtry
- Max-pooling: 5 max-pooling vrstev s 2x2 okny a krokem 2
- Plně propojené vrstvy: 3 vrstvy, z nichž první dvě mají 4096 kanálů a třetí 1000 kanálu pro klasifikaci do 1000 tříd
- Aktivační funkce: ReLU, je použita ve všech vrstvách
- Výstup: Softmax vrstva pro klasifikaci

26. Popište model U-Net.

- Původně navržen pro segmentaci buněk tkání v mikroskopických snímcích
- Navržen tak, aby byl efektivní i při trénování na malých datasetech
- Používá augmentaci dat

Architektura:

- Symetrická architektura písmene U, která se skládá z kontrakční a expanzní cesty

27. Popište model ResNet.

- Základní model pro detekci a segmentaci obrazu
- CNN navržená k řešení problému degradace při trénování velmi hlubokých sítí

Architektura:

- K dispozici v různých hloubkách
- ResNet18, 34, 50, 101, 512 – číslo značí počet vrstev

28. Popište princip neuronových sítí typu GAN a jejich typické využití.

- Generativní protichůdné sítě
- Učení bez učitele
- Třída neuronových sítí, které se autonomně učí vzorce ve stupních datech, aby generovaly nové příklady připomínající původní datovou sadu

Aplikace:

- Syntéza obrazu
  - Generování vysoce realistických obrazů, pokročilé úpravy, manipulace s atributy
  - Reklama, počítačová grafika
- Generování obrázku k textu
  - Kombinace GAN s technikami zpracování přirozeného jazyka
  - Generování obrázků z textového popisu
  - Elektronické obchodování – přeměna popisu produktů do vizuální reprezentace
- Generování a predikce videa
  - Syntéza realistického poutavého videa
  - Detekce deepfake, odhalování manipulací
  - Vývoj videoher – realistické postavy, scény a animace
- Adaptace domény a přenos stylu
  - Transformace obrázků do různých uměleckých stylů, přeměna fotografie na malbu, přenos stylu mezi doménami
  - Uplatnění v designu a reklamě

29. Popište fungování a učení generátoru a detektoru v GAN.

Generátor:

- Vstup
    - Vektor náhodného šumu s normálním nebo rovnoměrným rozdělením
- Architektura
    - Vstupní vrstva: vektor náhodného šumu
    - Plně propojené vrstvy: upravují vektor pro další zpracování
    - Dávková normalizace: technika použita mezi vrstvami ke stabilizaci učení a normalizaci výstupu předchozí vrstvy
    - Aktivační funkce: ReLU, Leaky ReLU
    - Transponované konvoluční vrstvy: převzorkují vstup z předchozí vrstvy do vyšší prostorové dimenze
    - Reshaping Layers: přetvoří data do požadovaného výstupního formátu
    - Výstupní vrstva: využívá funkci aktivace tanh nebo sigmoid
- Výstup
    - Generovaný obraz nebo text
- Učení
    - Generátor aktualizuje váhy na základě zpětné vazby od diskriminátoru
- Možný problém
    - Kolaps: produkování omezené škály výstupů, přestože vstup je pestrý

Diskriminátor:

- Vstup
    - Vzorky skutečných a generovaných dat
- Výstup
    - Hodnota mezi 0 a 1 – pravděpodobnost
- Architektura
    - Konvoluční vrstvy: pomáhají při extrahování funkcí ze vstupních obrázků
    - Dávková normalizace: mezi vrstvami se používá ke stabilizaci učení normalizací vstupu do vrstvy
    - Aktivační funkce: ReLU
    - Poolovací vrstvy: pro postupné snižování dimenze dat
    - Plně propojené vrstvy: ke zpracování prvků extrahovaných konvolučními vrstvami, které vyvrcholí konečnou výstupní vrstvou
    - Výstupní vrstva: typicky jeden neuron s esovitou aktivační funkcí, výstupem je hodnota pravděpodobnosti
- Možný problém
    - Příliš silný diskriminátor
        - Rychle začne poskytovat jistý výstup pro jakýkoliv vstup, což ztěžuje generátoru učení
    - Slabý diskriminátor
        - Nedává generátoru smysluplnou zpětnou vazbu ke zlepšení

30. Vysvětlete spojitost GAN s teorií her.

- Jedná se o hru dvou hráčů – generátor a diskriminátor
- Obě sítě jsou trénovány současně
    - Pokud generátor oklame diskriminátor, diskriminátor se musí zlepšit a naopak
- Proces pokračuje, dokud generátor nevytváří data téměř nerozpoznatelná od skutečných

31. Popište rozdíly mezi GAN a konvolučními neuronovými sítěmi.

- Ztrátová funkce
  - CNN používají mnoho různých ztrátových funkcí, diskriminátor GAN vždy používá binární ztrátu křížové entropie
- Smyčky zpětné vazby
  - CNN neexistuje žádná zpětná vazba s jinou sítí během jejich trénování, diskriminátor GAN pracuje v tandemu s generátorem
- Funkce aktivace výstupu
  - diskriminátor GAN obvykle používá funkci aktivace sigmoid ve výstupní vrstvě, aby poskytl skóre pravděpodobnosti
- Hloubka a složitost
  - diskriminátor GAN je často jednodušší a mělčí než CNN

32. Jmenujte důležité typy sítí GAN.

- Vanilla GAN
  - Základní model
  - Generátor a diskriminátor, oba jsou postaveny pomocí vícevrstvých perceptronů, tj. nepoužívají konvoluční vrstvy
- Deep convolutional GAN – DCGAN
  - Integruje architektury CNN do GAN
- Conditional GAN – CGAN
  - Zavádí koncept podmíněnosti, což umožňuje cílené generování dat na základě specifické dodatečné informace
- CycleGan
  - Používá dva generátory a dva diskriminátory
- Super-resolution GANS – SRGAN
  - Se zaměřuji na upscaling obrázků na vysoké rozlišení

33. Popište hlavní princip transformátorů vidění (vision transformers) a jeho typické využití.

- Model hlubokého učení, který využívá mechanismus pozornosti a posuzuje význam každé části vstupních dat
- Aplikace
  - Detekce objektů, segmentace obrazu, klasifikace obrazu, generativní modely
- Obrazy reprezentovány jako sekvence a jsou předpovězeny štítky tříd pro obrázek, což umožňuje modelům naučit se strukturu obrázku nezávisle

34. Popište fungování mechanismu pozornosti v transformátoru vidění.

Výpočet dotazů, klíč hodnota:

- Každé slovo ve vstupní frekvenci je převedeno na tří vektory

    - Dotaz
    - Klíč
    - Hodnota
- Tyto vektory jsou získány pomocí lineárních transformací původních vektorových reprezentací slov

Výpočet skóre pozornosti:

- Pro každé slovo se vypočítá skóre pozornosti vůči všem ostatním slovům
- To se provádí pomocí skalárního součinu dotazu Q a klíče K
- Normalizace pomocí softmax funkce
- Výsledkem je sada váhových koeficientů, které určují relevanci slov

Vážený součet hodnot:

- Každé slovo je reprezentováno jako vážený součet hodnot všech slov v sekvenci
- Tento krok dovolí modelu zachytit kontextový vektor každého slova vzhledem k ostatním slovům

35. Popište rozdíly mezi transformátorem vidění a konvolučními neuronovými sítěmi.

- ViT může zpracovávat vstupy libovolné velikosti, aniž by bylo nutné nadále měnit návrh modelu
- ViT může objevit korelace mezi různými prvky vstupního obrázku
- ViT je výpočetně efektivnější, protože má méně parametrů než CNN

36. Popište aktuální trendy ve vývoji metod a aplikací strojového učení z obrazových dat.

Edge AI – zpracování dat v reálném čase:

- Přesun z cloudu na místní zařízení bez internetového připojení
- Aplikace
    - Autonomní vozidla, výroba, obchod

Multimodální AI:

- Kombinace obrazu, zvuku a textu
- Aplikace
    - Zdravotní péče, smart cities, osobní asistenti

Aplikace ViT:

- Analýza celého obrazu současně posouvá oblast rozpoznávání
- Aplikace
    - Zabezpečení, elektronický obchod, zemědělství

Syntetická data:

- Namísto spoléhání se na obrázky z reálného světa se model může trénovat na syntetických, AI-generovaných obrázcích
- Aplikace
    - Autonomní vozidla, zdravotní péče

1. Popište, čím se zabývá zpracování přirozeného jazyka.

- Oblast umělé inteligence zaměřené na interakci mezi počítači a lidským jazykem
- Transformace z nestrukturovaného textu na strukturovaný
- Snaha převzít informace z nestrukturovaného textu do programu nebo jinak strojově použitelné struktury
- Cílem je analyzovat, porozumět a později generovat a reagovat na lidský jazyk

2. Popište, čím se vyznačují symbolové (resp. lexikonové) metody zpracování přirozeného jazyka a jaké mají výhody a nevýhody.

- Ručně psaná pravidla pro jazykovou analýzu
- Používá formální gramatiky, pravidlové skripty, ontologie a ručně psaná pravidla
- Používané metody
    - Regulární výrazy, syntaktická analýza, sémantické stromy
- Výhody
    - Interpretovatelnost, přesnost na specifických úlohách
- Nevýhody
    - Složitá údržba, špatná škálovatelnost
- Turingův test a první chatboti

3. Popište, čím se vyznačují statistické metody zpracování přirozeného jazyka a jaké mají výhody a nevýhody.

- Přechod od pravidel k pravděpodobnostním modelům
- Místo ručně psaných pravidel využívá modely založené na pravděpodobnostech a statistikách získaných z rozsáhlých dat
- Rozvoj latentní sémantické analýzy a vektorových reprezentací slov
- Používané metody
    - N-gramové modely, skryté Markovovy modely, Naivní Bayesovské klasifikátory
- Výhody
    - Schopnost zobecňovat na různá data, lepší škálovatelnost
- Nevýhody
    - Závislost na velkých datových souborech, obtížnější interpretovatelnost
- WordNet, IBM modely pro strojový překlad

4. Popište, čím se vyznačují metody zpracování přirozeného jazyka založené na neuronových sítích a jaké mají výhody a nevýhody.

- Použití neuronových sítí pro NLP
- Používané metody:
    - Rekurentní neuronové sítě, dlouhá krátkodobá paměť, latentní sémantická analýza, latentní Dirichletova alokace, Gated Recurrent Units
- Výhody
    - Lepší reprezentace slov než u statistických metod
- Nevýhody

    - Stále omezené porozumění kontextu, nižší efektivita výpočtu

5. Popište, čím se vyznačují transformátorové metody zpracování přirozeného jazyka a jaké mají výhody a nevýhody.

- Využití hlubokých neuronových sítí a transformerů
- Výhody
    - Vyšší přesnost, lepší kontextová analýza
- Nevýhody
    - Vyšší výpočetní náročnost, potřeba rozsáhlých dat
- Word2Vec
    - Neuronová síť pro word embeding

6. Popište účel analýzy sentimentu a stručně popište proces od předzpracování textu po samotnou analýzu pomocí jedné vybrané metody (kromě velkých jazykových modelů).

Účel:

- Někdy také „dolování názorů“
- Cílem je strojově určit subjektivní pocit nebo emoci z daného textu
- Používá se pro:
    - Sociální sítě – reakce na produkty nebo události
    - Zákaznická podpora – hodnocení spokojenosti
    - Finanční sektor – predikce tržních trendů na základě zpráv a komentářů
    - Mediální analýza – pochopení veřejného mínění o určitých tématech
- Výzvy jsou:
    - Ironie a sarkasmus
    - Dvojznačnost, idiomy, metafory
    - Doménově specifické výrazy
    - Multimodální vstup

Proces:

- Tokenizace
- Normalizace textu
- Odstranění stop slov
- Stematizace/Lemmatizace
- Analýza sentimentu

Stematizace:

- Stemmer vrací kořen/kmen slova
- Odstraňuje předpony a koncovky
- Problémy
    - Stejný kmen pro slova s různými významy
    - Různé kmeny slova pro slova se stejným významem

7. Popište tokenizaci a jak se používá.

- Rozdělení textu na menší jednotky – tokeny
- Základní krok nutný pro mnoho dalších úloh NLP
- Token může být slovo, věta, nebo i znak dle potřeby
- V náročnějších jazycích nemusí být mezery mezi slovy
- Moderní modely používají tokeny menší než slovo
    - Přirozený jazyk → při roze ný jazyk

8. Popište, co je stematizace a lemmatizace, jak se liší a kde se používají.

- Různé tvary slov normalizuje na jeden určitý tvar, se kterými si navazující metoda snáze poradí
- Závislé na jazyce textu

Stematizace:

- Vrací kořen/kmen slova
- Odstraňuje předpony a koncovky

Lemmatizace:

- Vrací tzv. lemmu – základní tvar slova
- Může doplňovat další informace o slově – druh, rod, číslo, pád
- Důležité pro fulltext vyhledávání
- Řeší problémy u stematizace
- Stále možné problémy s mnohoznačností a idiomy, pokud správně nepozná kontext

9. Popište vektorizaci, uveďte a vysvětlete princip alespoň dvou technik využívaných k vektorizaci.

- Převod textových dat na číselnou reprezentaci
- Číselná reprezentace slov je pro metody NLP lépe uchopitelná, méně paměťově i výpočetně náročná
- Různé metody do vektoru kódují různé informace
    - Přítomnost/nepřítomnost slova v dokumentu
    - Četnost slov v dokumentu
    - Pořadí slov
    - Kontext slov v rámci dokumentu
    - Význam

One-hot:

- Pro každé unikátní slovo 1 dimenze – obří vektory
- Binární hodnota pak určí (ne)přítomnost daného slova v dokumentu

Bag of Words:

- Tvoří slovník na základě obsahu dokumentů

- Uchovává četnost slova v dokumentu

N-gramy:

- Stejný princip jako Bag of Words, ale vektorizují více slov najednou
- Bigram = 2 slova, Trigram = 3 slova
- Dokáže zachytit pořadí slov

Tf-idf:

- Stejný princip jako Bag of Words
- Ubírá váhu slovům, která se vyskytují obecně bez ohledu na třídu, do které má dokument patřit

Word embedding:

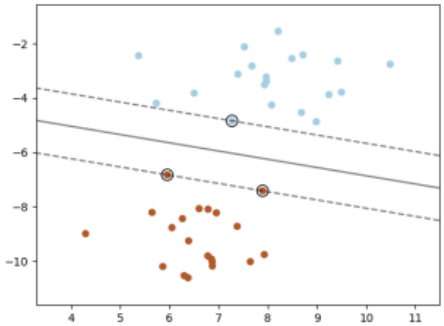
- Vektory vyjadřují význam slova – je potřeba více dimenzí

10. Popište naivní Bayesovský klasifikátor a jeho princip.

- Klasifikační metoda založená na Bayesově větě
  - Jak souvisí podmíněná pravděpodobnost nějakého jevu s otočenou podmíněnou pravděpodobností
- Naivní, protože předpokládá nezávislost všech atributů
  - To často není pravda, ale i tak metoda vede k dobrým výsledkům
- Počítá pravděpodobnost určité třídy Ck na základě pozorovaných dat X
- Výhody
  - Rychlý a výpočetně nenáročný
  - Nenáročný na množství testovacích dat
  - Snadno interpretovatelný
- Nevýhody
  - Naivní předpoklad nezávislosti může být v praxi porušen, snižuje se přesnost
  - Pokud je některá pravděpodobnost nulová, může to způsobit problémy
  - Horší výkon u složitějších vzorů, kde atributy nejsou závislé

11. Popište metodu podpůrných vektorů a její princip.

- Klasifikační metoda
- Hledání optimální hranice mezi třídami, která maximalizuje vzdálenost mezi nejbližšími body obou tříd a touto hranicí
- Pro lineární případ je hranice rovina



- Neoddělitelné třídy – ale stále lineární
    - Musíme povolit nějakou toleranci chyb
    - Měkká hranice
    - Nastavuje parametr C
- Neoddělitelné třídy – nelineární
    - Když pro oddělení nestačí lineární hranice
    - Výběr jiného kernelu umožní transformaci dat o dimenzi výš, tak, aby rovina oddělila třídy

12. Popište rekurentní neuronové sítě a jejich princip.

- Navrženy pro práci se sekvenčními daty – text, řeč, číselné řady
- Postupují sekvenčně jako klasické NN, kde jsou vstupy nezávislé, ale díky zpětné smyčce dokáže uchovávat stav skrytých neuronů
- Při trénování může docházet ke „zmizelému gradientu“
    - Po mnohonásobném násobení vah u prvních vrstev dojde k exponenciálnímu zmenšení oproti vahám posledních vrstev
    - Výsledkem je nestabilita, zpomalení, nebo úplné zastavení učení
- Dlouhá krátkodobá paměť
    - Řeší problém zmizelého gradientu
    - Umožňuje informaci překlenout i miliony iterací
    - Vstupní, výstupní a zapomínající brány rozhodují, co uchovat a co zahodit
    - Pro NLP využití umožňují pojmout delší kontext
- Využití
    - Strojový překlad, klasifikace, rozpoznání řeči
- Výhody
    - Schopnost modelovat sekvence
    - Při využití LSTM také dlouhodobé závislosti
- Nevýhody
    - Zdlouhavý trénink
    - Buď problém zmizelého gradientu pro standardní RNN
    - Nebo ještě vyšší výpočetní nároky pro LSTM

13. Vysvětlete problém zmizelého gradientu a jak je možné mu předejít.

- Při trénování může docházet ke „zmizelému gradientu“
    - Po mnohonásobném násobení vah u prvních vrstev dojde k exponenciálnímu zmenšení oproti vahám posledních vrstev
    - Výsledkem je nestabilita, zpomalení, nebo úplné zastavení učení
- Dlouhá krátkodobá paměť
    - Řeší problém zmizelého gradientu
    - Umožňuje informaci překlenout i miliony iterací
    - Vstupní, výstupní a zapomínající brány rozhodují, co uchovat a co zahodit
    - Pro NLP využití umožňují pojmout delší kontext

14. Stručně popište proces zpracování textu velkým jazykovým modelem (LLM).

- Tokenizace a případně primitívní vektorizace
- Word embedding
- Attention vrstva
    - Upravení významu slova podle kontextu
- Mzltilayer perceptron vrstva
    - Neuronová síť
- N opakování vrstev Attention a Perceptron
- Unembedding posledního vektoru
- Výpočet pravděpodobností slov pomocí funkce softmax a výběr jednoho následujícího slova

15. Popište embedding a jaký je jeho cíl.

- Transformace textových dat na vektory o stovkách až tisících dimenzí
- Blízké vektory = významově blízká data

Word embedding:

- Používá neuronové sítě nebo naučené matice k přiřazení hodnot vektoru každému slovu
    - První metody ještě bez LLN zohledňují kontext slova, ale různé významy „zprůměrují“ do jednoho vektoru
- Transformátory se self-attention mechanismem umožňují vzít v úvahu celý kontext a okolí slova
    - Kontextový embedding
- Vektory vyjadřující význam slova umožňující aritmetiku se slovy
- Podobnost měřená pomocí cosinové vzdálenosti

16. Vymyslete alespoň dva jednoduché příklady aritmetiky se slovy, kterou umožňuje word embedding.

- Tento klasický příklad ukazuje, že když od vektoru slova „král“ odečteme pojem „muž“ a přičteme „žena“, získáme vektor velmi blízký slovu „královna“

vec("král") - vec("muž") + vec("žena") ≈ vec("královna")

- Tento příklad ukazuje geografické nebo geopolitické vztahy: hlavní město státu. Pokud odečteme Francii od Paříže a přičteme Itálii, získáme město, které k Itálii plní stejnou roli – Řím

vec("Paříž") - vec("Francie") + vec("Itálie") ≈ vec("Řím")

17. Popište, co je velikost kontextu a jak ovlivňuje využívání velkého jazykového modelu (LLM).

- Určuje kolik slov najednou je model schopný vzít na vstup

- Méně tokenů → vyplní se prázdnými tokeny
- Více tokenů → buď na etapy, nebo se prostě vezmou jen poslední → u chatbota zapomínání na kontext probíraný dříve

18. Popište, na jakém principu pracuje word embedding ve velkém jazykovém modelu (LLM) a co je jeho vstupem a výstupem.

- Každý model má své vlastní parametry pro embedding – embeddin matice We
- Začíná náhodně, trénování na datech se nastaví
- Každé slovo je pomocí embedding matice zakódováno do vektoru s N hodnotami, kde N je počet dimenzí, se kterými daný model pracuje
- Kolik slov zná záleží na velikosti slovníku tokenů
- We má rozměry počet_dimenzí * velikost_slovníku

Princip:

- Ze slovníku a tokenů je nejprve vytvořena jednoduchá binární matice
- Každé slovo je označeno 1 jako při one-hot
- Pak se násobí maticí We

$$\underset{\text{velikost\_kontextu}}{\begin{bmatrix} 0 & 1 & \cdots & 0 \\ 1 & 0 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}}_{\text{velikost\_slovníku}} \times [W_E] = \underset{\text{velikost\_kontextu}}{\begin{bmatrix} 8{,}6 & -0{,}1 & \cdots & 2{,}5 \\ 1 & -4{,}9 & & -1 \\ \vdots & & \ddots & \vdots \\ -0{,}5 & 2{,}3 & \cdots & 7{,}1 \end{bmatrix}}_{\text{počet\_dimenzí}}$$

- Po embeddingu se k vektorům ještě přičtou hodnoty vyjadřující pozici slov
    - Jinak ale slovo zatím zůstává bez kontextu okolí
- Poté prochází několika iteracemi střídavě
    - Self-attention vrstvy
    - Peceptron vrstvy
- Poté je potřeba rozklíčovat výstup na další následující slovo – token

19. Popište, jak ve velkém jazykovém modelu (LLM) probíhá unembedding a jaký výstup z něj získáme.

- Výstupem z Attention a Perceptron vrstev je matice o rozměrech velikost_kontextu * počet_dimenzí
- Tzn. Kolekce vektorů ve stejném počtu jako na začátku, kde každý vektor byl upraven tak, aby obsahoval předpověď pro následující slovo pro danou pozici
- Pro generování slov nás zajímá jen poslední vektor, který musíme přeložit zpátky na slovo – token
- Stejně jako na začátku přeložily vektor pomocí embedding matice We, podobná matice Wu je využita i pro překlad zpátky na tokeny
- Některé modely používají stejnou matici, jen transponovanou

Princip:

- Vypočítané vektory se neshodují přesně se slovníkem, takže výsledkem není jasné označení slova, ale spíš ohodnocení všech slov ve slovníku
- Toto hodnocení si můžeme představit jako porovnávání, jak blízko jsou jednotlivá slova k významu, který jsme získali průchodem LLM
- Vektor ohodnocení v tuto chvíli obsahuje i záporné hodnoty a také hodnoty větší než 1

20. Vysvětlete alespoň jednu metodu používanou po unembeddingu ve velkém jazykovém modelu (LLM) k ovlivnění výběru finálního slova či tokenu.

Softmax

- K výpočtům uvnitř transformátoru a na výstup pro pravděpodobnosti následujícího slova
- Transformuje hodnoty tak, aby byly mezi 0 a 1 a zároveň jejich suma byla 1
- K ovlivnění „kreativity“ modelu se využívá parametr Temperature

Top-K

- Podle vektoru zpracovaného pomocí softmaxu vybere k nejvyšších ohodnocení
- Z toho výběru se nakonec vybere a použije finální následující slovo
- Pro k=1 model vybere vždy hodnotu s nejvyšší pravděpodobností

21. Vysvětlete, jak a v jaké fázi procesu ovlivňuje parametr temperature výstup velkého jazykového modelu (LLM).

- Během aplikace softmax
- Ovlivňuje výpočet pravděpodobností na výstupu
- Nízká hodnota
    - Model vybírá spíše slova, kterými si je jistý
- Vysoká hodnota
    - Model vybírá ze širší palety, ale více riskuje, že se netrefí – bude halucinovat

22. Vysvětlete princip a účel attention vrstvy ve velkém jazykovém modelu (LLM).

- Umožňuje modelu dynamicky rozhodnout, kterým částem vstupního textu má věnovat největší pozornost
- Pro každý token se počítá vážený průměr ostatních tokenů na základě relevance, kterou určuje attention score
- Účel
    - Zachytit kontext
    - Porozumění významu
    - Flexibilita

23. Vysvětlete, jak a v jaké fázi zpracování textu velkým jazykovým modelem (LLM) se mění význam slova či tokenu na základě kontextu okolních slov.

- Význam slova se neinterpretuje pevně, ale dynamicky se mění v závislosti na kontextu okolních slov
- Model při každé pozici „zvažuje“ jaké informace z okolních slov jsou pro daný tok důležité
- To se dělá pomocí
    - Attention mechanismu
        - Každý token získá kontextualizovanou reprezentaci na základě ostatních tokenů
    - Transformačních vrstev
        - Každá vrstva upravuje vektor tokenu tak, aby odrážel stále hlubší porozumění celkovému kontextu
- V jaké fázi
    - Po embedování
        - Každý token je převeden na fixní vektor, který nemá kontext
    - V průběhu attention vrstev
        - V každé vrstvě mechanismus „míchá“ informace mezi tokeny podle jejich relevance
    - Po několika vrstvách
        - Každý token má nový vektor, který už neznamená jen samotné slovo, ale jeho význam v dané větě

24. Vysvětlete princip a účel perceptronové vrstvy ve velkém jazykovém modelu (LLM).

- Nachází se za attention vrstvou
- Zpracování informací na úrovni jednotlivých tokenů
    - FNN pracuje s každým tokenem zvlášť a umožňuje složitější transformaci významu, než by zvládla samotná attention vrstva
- Zvyšuje výpočetní kapacitu modelu
    - Díky nelinearitám umožňuje model učit se složitější vzory
- Posiluje kontextualizaci
    - I když FNN nepracuje s tokeny přímo, pomáhá modelu jemně doladit význam každého tokenu na základě jeho kontextem upraveného stavu z attention

25. Vysvětlete, jak a v jaké fázi zpracování textu velkým jazykovým modelem (LLM) se do generovaného textu promítají informace a znalosti získané při trénování modelu.

- LMM získává při trénování statistické znalosti o jazyce, světě, faktech a vztazích mezi slovy
- Znalosti se neukládají do databáze, ale rozprostřou se do vah neuronové sítě
    - Ty určují, jak model reaguje na různá vstupní slova a kombinace

- Váhy určují
    - Jaký význam přisoudit slovům na základě kontextu
    - Jaké informace a souvislosti nabídnout
    - Jaká slova navrhnout jako nejpravděpodobnější pokračování textu

Fáze:

- Při zpracování vstupu
    - Model získá vektorovou reprezentaci slov, ale samotné „vědění“ ještě nevyužívá
- Během průchodů skrz vrstvy modelu
    - Právě zde se naplno uplatňují naučené váhy
    - Ty určují, jak model interpretuje význam slov, jak reaguje na kontext
- Při výběru dalšího tokenu
    - Na základě váhových parametrů model určí pravděpodobnosti dalších slov, a tím vybere další token do textu

26. Vysvětlete, co pro velký jazykový model znamená počet dimenzí vektorů a jak ovlivňuje využívání modelu.

- V LLM jsou slova a tokeny reprezentovány jako vektory – číselné řady
- Počet dimenzí udává, kolik čísel je ve vektorové reprezentaci jednoho tokenu
- Tento styl se týká:
    - Embedding vektorů
    - Skrytých stavů v jednotlivých vrstvách transformeru
    - A ovlivňuje celý výpočetní tok modelu

Ovlivňuje:

- Výpočetní náročnost
    - Více dimenzí = vyšší přesnost, ale větší spotřeba paměti a zdrojů
    - Větší dimenze = vetší matice = náročnější operace
- Kapacita modelu
    - Vyšší počet dimenzí umožňuje zachytit složitější vztahy a jemnější významové nuance
    - Model s malým počtem dimenzí může být rychlejší, ale méně chytrý
- Velikost modelu a nasazení
    - Modely s vyššími dimenzemi bývají většinou hůře se nasazující na slabší hardware
    - Menší dimenze jsou vhodné pro rychlou inferenci nebo běh na klientském zařízení

27. Popište, co je Langchain a jaký je jeho účel.

- Framework pro vytváření aplikací využívajících jazykové modely
- Umožňuje propojovat LLM s externími nástroji
- Usnadňuje tvorbu řetězců, které kombinují více kroků zpracování
- Podporuje paměť, práci s dokumenty a agentový přístup

28. Vysvětlete, jak ve velkém jazykovém modelu (LLM) funguje paměť a jak ji lze řešit v chatovacích aplikacích.

- LLM je stateless – nemá paměť
- Kromě procesu učení, nezanáší do parametrů nové informace → nepamatuje si předchozí dotazy, pokud nejsou obsažené v aktuálním dotazu
- Chatovací LLM si v každém dotazu nesou i předchozí zprávy v konverzaci – alespoň dokud jim stačí velikost kontextu
- Při volání LLM v python musíme do každého následujícího dotazu přiložit všechny předchozí

29. Vysvětlete, co je RAG a k čemu se využívá.

- Retreival-Augmented Generation
- LLM má k dispozici ještě jiný zdroj než jen samotný dotaz

Jak funguje:

- Dotaz přijde do systému
- Retreiver najde relevantní texty nebo dokumenty z databáze / znalostní báze
- Generátor použije vstupní dotaz + nalezené informace a vygeneruje odpověď

Slouží k:

- Odpovídání na dotazy nad vlastními daty
- Zlepšení přesnosti odpovědi
- Aktualizovatelnost
- Zachování faktické správnosti

30. Vysvětlete, jak můžeme dát velkému jazykovému modelu (LLM) k dispozici velké množství textu či dokumentů, aniž bychom byli omezeni velikostí kontextu.

- LLM mají omezenou délku kontextu – např. 4000 tokenů
- Pokud chceme pracovat s větším objemem dat, používají se techniky, které umožňují přístup k dokumentů bez nutnosti je přímo „nacpat“ do vstupu modelu

RAG:

- Nejčastější přístup
- Všechny dokumenty se předem rozkouskují a uloží jako vektorové reprezentace
- Když přijde dotaz, pomocí vektorových vyhledávání se najdou relevantní části
- Tyto části se vloží do promptu spolu s dotazem → šetří kontext, ale využívá velké množství dat

External memory:

- Model se propojí s databází nebo systémem, který uchovává informace
    - Položí dotaz do DB

    - Získá výsledek, který pak dále zpracuje a použije ve výstupu
- Tím se velké množství dat neukládá do kontextu, ale dotahuje se v reálném čase

Chaining:

- Místo jednoho dotazu se konverzace rozkouskuje do více menších kroků

Hierarchické zpracování:

- Text se nejprve zpracuje po částech
- Každá část se shrne → tyto shrnuté části se pak opět shrnou na vyšší úrovni
- Tím vznikne kompaktní přehled, který se vejde do kontextového okna

31. Popište, jak funguje databáze vektorů (vector store) a k čemu ji lze využít.

- Speciální typ databáze, která neukládá klasická data jako text a tabulky, ale číselné vektory → reprezentace slov, vět, dokumentů

Účel:

- Umožnit rychlé a efektivní hledání podobných významů, nikoliv jen přesnou shodu

Jak funguje:

- Vytvoří embedding
    - Textový dokument se převede na embedding – číselný vektor
- Uložení vektorů do databáze
    - Každý embedding se uloží spolu s původním textem do vektorové DB
- Vyhledávání dotazu
    - Když uživatel zadá dotaz, převede se na embedding
    - Pak se porovná s uloženými vektory a vrátí ty, které jsou nejblíže – významově nejpodobnější

K čemu:

- RAG – klíčová součást
- Sémantické vyhledávání
- Chat s vlastními daty
- Doporučovací systémy
- Detekce duplicit / podobností

32. Popište, co umožňují agenti (z frameworku Langchain) pro aplikace s velkými jazykovými modely (LLM).

- Umožňují LLM provádět akce
    - Interagovat s wikipedií, provádět vyhledávání na internetu, kalkulačka…
- Mají k dispozici nástroje, které rozšiřují jejich moznosti

Typy:

- Zero shot ReAct

    - Pro jednotlivé interakce bez paměti
- Conversational ReAct
    - Pamatuje si předchozí dotazy
    - Ideální pro chatboty
- ReAct Docstore
    - Vyhledávání informací
    - Na wiki, v souboru…
- Další pokročilejší agenty nabízí rozšíření LangGraph

Vyhledávací agent

- Pokud nezná odpověď, sestaví dotaz pro vyhledávač
- Z nalezených dat si vybere, co potřebuje, a uloží k dalšímu zpracování
- Opakuje, dokud nemá všechny potřebné informace k zodpovězení dotazu uživatele

Příklady aplikací:

- Extrakce informací z faktur a vytvoření přehledu
- Vytvoření článku na zadané téma
- Generování obsahu na základě obrázku
- Zhodnocení životopisů ve složce
- Analýza dat z CSV
- Vytváření SQL dotazů z textového zadání
- Řešení matematických úloh