Univerzitet u Novom Sadu
Fakultet tehničkih nauka
Departman za računarstvo i automatiku
Katedra za primenjene računarske nauke

Osnovne akademske studije
Računarstvo i automatika

Napredne arhitekture informacionih sistema

Seminarski

Vektorske baze podataka kao podrška pretrazi slika

Član tima (TIM-13) :

- Student 1: *Vukašin Kalaba RA 44/2021*

Uvod

Slika je dosta teška za pretragu klasičnim putem, metapodaci slike su dosta ograničeni i ne mogu da pruže zadovoljavajuću pretragu. Ključne reči (metapodaci) koji su automatski ili ručno pridruženi nekoj slici često se susreću sa raznim problemima prilikom pretrage. Neki od njih su: nepotpun podatak, upotreba različitog jezika, komplikovanija slika (više objekata na slici) ...
Takođe oslanjanje na način pristupa poređenja slike piksel po piksel stvara mnoge nedostatke u procesu pretrage. I sama promena osvetljenja slike menja njene piksele, a samim tim i sličnost između dve slike, gde se jednoj od njih promeni osvetljenje, drastično opada. Slični problemi nastaju i sa rotacijom, kao i skaliranjem slike, dok će se dve iste slike sa različitim pozadinama gledati kao dve potpuno različite slike.

Stoga razvoj vektorskih baza podataka je dosta doprineo lakšoj pretrazi slika. Cilj je da se svaka slika vektorizuje (pretvori u određeni vektor), gde se kasnije kao takva čuva u vektorskoj bazi, nad kojom se vrše upiti koji koriste razne algoritme pretrage (kNN, ANN…). Slika se u vektorskoj bazi predstavlja kao embedding – niz brojeva koji poseduje semantičku suštinu slike. To čini poređenje slika daleko efikasnijim. Tu se onda javlja i prirodan primer primene vektorskih baza jer korisniku unosom slike, skice ili nekog teksta koji opisuje sliku, kao upit, sistem pronalazi slike koje su vektorski najbliže toj slici tj. traženje najbližih vektora u prostoru velike dimenzionalnosti.

Zbog toga je došlo do sve veće potrebe za primenom ovog pristupa prilikom pretrage slika, kao i potreba njegove implementacije u tehnologije svakodnevnog života. Neki od najpopularnijih načina primene su: u komercijalne svrhe, u društvenim mrežama, u medicinske svrhe … Svugde gde na osnovu jedne slike možemo da nađemo njoj slične.

Domen primene

Razvoj vektorskih baza doprineo je tome da sve više i više viđamo upotrebu tog pristupa u svakodnevnom životu.
Neki od glavnih primera upotrebe vektorskih baza kao podrška pretrazi slika su:

Komercijalne svrhe (E-commerce)

Pretraga slika u ovom scenariju funkcioniše tako što, korisnik uslika artikal koji želi da pronađe u nekoj od web prodavnica, sistem mu onda pretragom vrati slične proizvode (vektorski najbliže slici kojoj je poslao). Pretraga slike prati sledeće korake:

1. Ekstrakcija osobina iz slike (iz slike se izvlači embedding)
2. Vektorska pretraga (izvučeni embedding se šalje u vektorsku bazu, koja vraća top-K najbližih proizvoda)
3. Re-ranking + filteri (rezultati se dodatnoređaju i filtriraju)
4. Učenje iz klika (klikovi korisnika služe za fino podešavanje modela, i boljeg pogađanja namere

Jedan od najpoznatijih web prodavnica koje koriste ovaj pristup je Alibaba. Koristeći ovaj pristup njima se javljaju mnogobrojni problem, a jedan od njih je domenski jaz (stvarna slika koju korisnik šalje vs. katalog slika), masovno ažuriranje indeksa i kategorizacija uz jako ograničenje vremena. Da bi pretragu učinili što efikasnijom, a samim tim i ispravili nedostatke, oni kombinuju detekciju objekata, učenje reprezentacija, binarne/kvantizovane indekse tako postižu brži i tačniji odgovor, pri stalnom prilivu novih artikala.

Društvene mreže (preporuka vizuelnog sadržaja)

Korisnik kada otvori neku objavu, na osnovu te objave on dobija vizuelno slične ideje ili sadržaj koji bi njemu bio od nekog značaja (look-alike sadržaj). Realizaciju ovog pristupa prate sledeći koraci:

1. Embeddings za slike koje pune vektorsku bazu
2. On-the-fly pretraga – kada se gleda neki post (slika), sistem iz embeddinga slike koja predstavlja post izvuče najbliže komšije, a zatim personalizuje dalju pretragu na osnovu postova na koje je korisnik ulazio
3. Specifični vektori za stil (boje, teksture …) ili specifični modeli za domen (npr. moda) dopunjuju semantičku sličnost.

Pinterest i slične platforme godinama investiraju u vizuelnu pretragu i personalizaciju u ogromnom broju (milijarde pretraga mesečno), a vizuelna pretraga je sve više i više počela da preuzima primat nad standardnom tekstualnom pretragom, kada su u pitanju ideje i stilovi.

Bezbednost (prepoznavanje lica)

Pronalaženje svih pojava istog lica (verifikacija/identifikacija) u velikim foto-bazama. Ovaj pristup oblikuju sledeće stavke:

1. Specijalizovan embedding (FaceNet i njegovi naslednici) – gde su vektori koji čine slike istog lica blizu jedni drugih, dok su oni različiti daleko
2. Mala dimenzija vektora sa kombinacijom kosinusnog ili euklidskog pristupa za računanje rastojanja između dva vektora uz ANN indeks pretragu (brza identifikacija nad velikim skupovima).
3. Izazovi – osvetljenje, poza, „look-alike“ osobe (zahteva veliki trening, i verifikacione pragove)

Ovaj pristup je jedna od najklasičnijih, i najuspešnijih primena embedding + najbliži sused paradigme.

Medicina (pronalaženje sličnih snimaka)

Lepo je videti da se vektorske baze mogu primenjivati i u humane svrhe. U ovom pristupu radiolog može iz priloženog snimka nekog pacijenta da istraži slične slučajeve koje je ta klinika imala i na osnovu toga lakše uspostavi dijagnozu.

1. Specifičan feature extractor (na primer DenseNet podešen na medicinskim podacima) koji generiše embedding
2. Vektorska pretraga – vraća slične snimke
3. Evaluacija – preciznost na nekom uzorku od k rezultata i klinička relevantnost bitniji su od sirove klasifikacione tačnosti. Cilj je podržati odluku i ubrzati poređenje slučajeva.

Iako nije još zastupljen u medicini, radovi na ovu temu pokazuju da se ovakvi sistemi mogu uklopiti u realne tokove rada, uz poboljšanja u top-K (preciznost na uzorku od K), relevantnosti i brzini.

Implementacija

Proces izgradnje sistema za pretragu slika zasnovanog na vektorskim bazama uvodi sledeće etape:

1. Priprema podataka (osnova svakog sistema)

- Skup slika: treba sakupiti reprezentativne slike za svoj domen (npr. e-commerce proizvodi, arhiva fotografija, medicinski snimci). Ako imamo metapodatke (kategorija, brend, autor, datum), treba i njih sačuvati – biće nam od značaja za filtriranje i re-ranking.
- Preprocesiranje: resize (npr. 224×224), normalizacija po statistici modela (mean/standardna devijacija) i pažnja na orijentaciju (EXIF - Exchangeable Image File Format).
- Kvalitet: najbolje bi bilo izbaciti duplikate ili ekstremno loše slike (previše male ili mutne) jer kvare embedding prostor i metrike.

2. Izbor modela za embedding (srce sistema)

- Modeli:
  - CNN (ResNet/VGG/EfficientNet): provereni, brzi, dobri kao „backbone“ za domenske probleme.
  - Vision Transformers (ViT): bolje hvataju globalni kontekst, dok zahtevaju više podataka ili pretreniranje
  - CLIP – tip multimodalni model: omogućava tekst → slika upite, bez doterivanja za svaku klasu
- Dimenzionalnost i normalizacija: vektori dimenzija 256 – 1024 su česti, L2 – normalizacija (euklidska norma za računanje razdaljine između vektora) ili unit – norm (norma vektora je 1) pomaže kod kosinusne sličnosti, a po potrebi se koriste i PCA (Principal Component Analysis – linearno smanjenje dimenzionalnosti) i PQ (product quantization – standardna tehnika u Facebook AI Similarity Search-u za smanjenje memorije) za kompresiju i bržu pretragu. Bitno je da se ista obrada (normalizacija/PCA) primeni i pri izvršavanju upita, i pri indeksiranju.
- Transfer learning i domen: u medicini ili specijalizovanim domenima često je isplativ fine-tuning na manjem, ali reprezentativnom skupu.

3. Indeksiranje i izbor vektorske baze

- HNSW (Hierarchical Navigable Small World – grafski baziran algoritam za ANN pretragu): odlična tačnost i brzina pretrage, idealan do nekoliko miliona vektora. Parametri tipa efConstruction, M (gradnja) i efSearch (upit) omogućavaju nam fino podešavanje.
- IVF (Inverted File Index) + opcioni PQ (kvantizacija): podeli prostor u **nlist** klastera, pri upitu pretraži samo **nprobe** najbližih klastera. Sa PQ drastično se smanjuje memorija (uz malu kaznu na tačnosti). Ovo je dobar put kada treba da imamo na desetine ili stotine miliona slika.
- On-disk varijante: ako podaci ne staju u RAM, postoje vektorske baze podataka koje drže vektore i indekse na disku uz razumno vreme odziva (latencija).

Preporuke:

- do 5M vektora – savet je da krenemo sa HNSW (jednostavnije, brzo, često najbolji izbor).
- 10M i više – IVF ili IVF+PQ (štedi nam dosta RAM, i drži vreme odziva pod kontrolom).
- Neophodno je uvek testirati na svom domenu jer ne postoji univerzalno najbolji.

4. Arhitektura sistema (servis za generisanje embeddinga)

- Offline (ingest): prolazi se kroz veliku kolekciju, pretvaranje slike u vektore i upisivanje u vektorsku bazu (i/ili objekat skladište + meta baza). Obrada u serijama (batch), uz paralelizaciju – umesto da svaku sliku obrađujemo pojedinačno, grupišemo ih u veće pakete i puštamo kroz model u jednom prolazu, više radnika/servisa radi paralelno nad različitim serijama.
- Online (u trenutku upita): prima se slika ili tekst od korisnika, kroz isti model dobija se embedding upita (isti način kao kod offline), i vrši se slanje na pretragu.

Dva važna saveta:

Poželjno je verzionisati modele tako što se čuva verzija modela uz svaki vektor, ukoliko bude došlo do menjanja modela zna se koji embedding je kojim modelom nastao.

Poželjno je keširati popularne upite, i njihove embeddinge, time štedimo milisekunde.

5. Metapodaci i hibridna pretraga

Vektorska pretraga nam vraća kandidate (ID-jeve). Međutim korisnik želi i određene filtere: kategorija, brend, cena, datum …

Zato se u praksi upotrebljava hibridni pristup:

- pre-filter: prvo se kolekcija suzi običnim upitom (SQL/NoSQL), pa ANN samo nad preostalim vektorima
- post-filter/re-ranking: prvo ANN na celoj kolekciji vektora, pa se rezultati preslože i filtriraju po metapodacima

6. Evaluacija: kako znamo da radi dobro?

- Kvalitet:
  - *precision@k* (broj relevantnih rezultata u prvih k) / k
  - *recall@k* (broj relevantnih u prvih k) / (ukupan broj relevantnih u kolekciji)
  - *mAP (mean Average Precision) / NDCG (Normalized Discounted Cumulative Gain)* – (ako nam je potrebna fina evaluacija po rejtingu)
- Performanse:
  - latencija P95 i P99 (koliko brzo sistem odgovara za većinu zahteva, 95 i 99 predstavljaju procente zahteva za koje korisnik dobija odgovor u nekom zadatom vremenskom intervalu)
  - memorijski otisak (koliko RAM-a nam zauzme indeks)
  - trošak izgradnje / reindeksiranja (koliko traje i da li se može raditi u hodu)

Saveti za postavke:

- HNSW: podizanjem efSearch-a dobijamo bolju tačnost (skuplje po latenciju)
- IVF: povećanjem nprobe (pretražujemo više klastera, što je tačnije ali sporije), podešavanjem nlist (dobijamo više klastera, gde dobijamo finiju podelu)
- PQ: više kodnih knjiga/bitova prouzrokuje manji gubitak kvaliteta, ali više memorije

7. Izvršavanje upita

1. Ulaz: korisnik pošalje sliku (ili tekst)
2. Embedding: isti model/obrada kao u ingestu – dobija se vektor upita
3. ANN pretraga: traži se top-K najbližih vektora (računanje kosinusne sličnosti i euklidskog rastojanja)
4. Re-ranking: kombinovanje semantičke sličnosti sa metapodacima (npr. prioritizuj novije, prioritizuj one koji su na stanju, prioritizuj relevantne po kategoriji…)
5. Uklanjanje duplikata i raznolikost: cilj je da se uklone duplikati i da se rezultatima da malo na raznolikosti, da vraćena lista ne bi bila sačinjena od istih slika
6. Odgovor: vrati se top-K sa naslovom, sličicom, linkom, cenom … Šta god da domen traži

Ograničenja, etička i pravna pitanja

Vektorske baze i sama pretraga slika pomoću njih nam donose razne prednosti, međutim bitno je napomenuti i da upotreba ovog načina pristupa dolazi sa nizom ograničenja i odgovornosti. Neki od njih su:

- Zavisnost od podataka i pristrasnost: Podaci na kojima je model učio direktno utiču na kvalitet pretrage, npr. ako su trenirani skupovi neuravnoteženi (premalo slika određenih grupa, uzrasta, boja kože), embedding prostor može biti pristrasan.
- Objašnjivost: Često se zbog kompleksnosti sistema koji je koncipiran na dubokim embedding-ovima ne može objasniti zašto su baš te slike ušle u top-K. Nedostatak objašnjivosti utiče na poverenje korisnika i internu validaciju posebno u oblastima kao što je medicina.
- Operativni troškovi i latencija: Rast broja slika (kolekcije) – na desetine miliona slika, podiže troškove skladištenja i održavanja indeksa. PCA, PQ, on-disk pretraga smanjuju troškove, ali u istu meru mogu spustiti tačnost.
- Privatnost i očekivanja korisnika: Korisnici često ne žele da se njihov sadržaj koristi za pretragu sličnosti van prvobitne svrhe. Pretraživanje po sličnosti može olakšati profilisanje, praćenje i neželjene identifikacije (npr. ako se na nekoj fotografiji nađete u pozadini a pritom niste želeli da vas neko fotografiše, već je tako slučajno ispalo)
- Biometrija i osetljive oblasti: Prepoznavanje lica i druge biometrijske tehnike (otisak prsta ...) spadaju u posebno osetljive stvari, gde je nedopustivo napraviti grešku. Stoga se u ovom domenu kao pomoć obavezno koriste pragovi i ljudsko nadgledanje. Zbog osetljivosti same oblasti bitno je navesti i to da se od korisnika ovog pristupa pretrage slika očekuje da isti ne zloupotrebe kako bi nekom naškodili.
- Pristrasnost modela: Ako sistem sistematski plasira sadržaj određenih grupa ili pogrešno rangira rezultate, on produbljuje postojeće nejednakosti. Potrebno je planirati i sprovoditi periodične audite pristrasnosti, kao i mehanizme za žalbe i ispravke.
- Kontrola i transparentnost: Korisnici bi trebalo da budu informisani da sistem koristi pretragu zasnovanu na sličnosti i da imaju mogućnost da isključe svoje sadržaje iz takvog indeksiranja kada je to moguće.

Zaključak

Za pretragu slika, vektorske baze su logičan izbor jer rade ono što je ključno za datu pretragu: brzo pronalaze semantički najbliže primere u ogromnim kolekcijama slika. Razlika između uspešnog i prosečnog sistema najčešće nije u jednoj arhitekturi, nego u pažljivom i tačnom spajanju više njih: dobar embedding za odgovarajući domen, pravi indeks i njegovo fino podešavanje, kao i zdrava operativa (ingest, ažuriranje, evaluacija). Kada na to dodamo multimodalnost (CLIP), dobijamo sistem koji odgovara i na upit u formi rečenice (šta želimo da pronađemo u bazi) i na upit gde prosledimo sliku nečega i kao odgovor dobijemo slične slike.

Vektorska pretraga slika donosi moćne mogućnosti, ali zahteva odgovorno i detaljno planiranje. Ograničenja (pristrasnost, objašnjivost …), kao i razna etička pitanja (privatnost, pravičnost, …) i pravne obaveze (osnov obrade, posebne kategorije podataka, autorska prava …) moraju biti planirana i analizirana od prvog dana. Kada se ova pitanja tretiraju temeljno, postupno i sistematički – kroz pravilne politike, tehničke mere i samu transparentnost – sistem koji smo kreirali daje najbolje rezultate, bez neželjenih posledica po korisnike i organizaciju.

Literatura (linkovi)

- FAISS i ANN indeksi (osnove i PQ/OPQ, skaliranje): „The Faiss Library“. arXiv
- Pregled vektorskih baza i ANN pristupa (hash/tree/graph/quantization): A Comprehensive Survey on Vector Database. vector_database
- Milvus (IVF/IVF_PQ/HNSW/SCANN) i detaljna objašnjenja, dokumentacija. Milvus
- Weaviate (HNSW – karakteristike i trade-off): dokumentacija. Weaviate Documentation
- Qdrant (cosine/dot, on-disk HNSW, tuning kvaliteta): dokumentacija. Qdrant
- E-commerce „na skali“: Visual Search at Alibaba (sistem, izazovi, binarni/kvantizovani indeksi, učenje iz klika). Visual Search
- Društvene mreže / discovery: Pinterest engineering (evolucija vizuelne pretrage) i noviji AI trendovi. Medium
- Medicina (CBIR sa DenseNet + FAISS, BIRADS): arXiv rad. Medicina
- Near-duplicate/copy-detection (pregled metoda i izazovi skale): pregledni rad. duplicate detection