



SISTEM ZA PRETRAGU SLIKA ZASNOVAN NA CLIP MODELU I VEKTORSKIM BAZAMA

IMAGE SEARCH SYSTEM BASED ON THE CLIP MODEL AND VECTOR DATABASES

Katarina Artukov, *Fakultet tehničkih nauka, Novi Sad*

Oblast – PRIMENJENE RAČUNARSKE NAUKE I INFORMATIKA

Kratak sadržaj – *Ovim radom predstavljen je sistem za pretragu slika zasnovan na CLIP modelu i vektorskim bazama podataka, koji omogućava efikasnu pretragu vizuelnog sadržaja na osnovu tekstualnih ili slikovnih upita.*

Ključne reči: *Pretraga slika, vektorske baze, multimodalna ugnježđenja, mašinsko učenje, CLIP model.*

Abstract – *This thesis presents an image search system based on the CLIP model and vector databases, enabling efficient search of visual content using textual or image-based queries.*

Keywords: *Image search, vector databases, multimodal embeddings, machine learning, CLIP model.*

1. UVOD

U doba kada količina digitalnih slika raste neviđenom brzinom proizilazi potreba za preciznom i efikasnom pretragom vizuelnog sadržaja. Uspešan sistem za pretragu mora biti u stanju da protumači korisnikove upite, zaključi njihove namere, a zatim primeni strategiju pretrage koja će verovatno vratiti relevantne rezultate. Tradicionalni sistemi za pretragu često zavise od tekstualnih oznaka (tagova) ili metapodataka koji se ručno dodaju, što nije održivo rešenje za velike baze podataka i često nije dovoljno tačno u reprezentaciji sadržaja slika.

Razvoj pretraživača slika uz korišćenje savremenih tehnika mašinskog učenja i obrade podataka omogućava nove načine pristupa informacijama i sadržaju, olakšavajući pronalaženje vizuelno sličnih ili konceptualno povezanih slika.

U ovom radu, razvijen je sistem za pretragu slika koji se oslanja na CLIP (*Contrastive Language–Image Pretraining*) model za generisanje multimodalnih ugnježđenja [1], koja kombinuju tekstualne i vizuelne karakteristike u jedinstveni prostor vektora. Ugnježđenja se zatim indeksiraju u *Pinecone* vektorsku bazu podataka.

NAPOMENA:

Ovaj rad proistekao je iz master rada čiji mentor je bio dr Dragan Ivanović, red. prof.

Integracija sa *Pinecone* vektorskom bazom [2] pruža skalabilno i brzo rešenje za pretragu i indeksiranje velikog broja slika. S obzirom na mogućnost indeksiranja multimodalnih ugnježđenja i brzog pretraživanja, ovaj sistem postaje izuzetno koristan u poljima kao što su istraživanje i pronalaženje sličnog vizuelnog sadržaja, marketing, analiza društvenih medija i kreativne industrije. Osim što unapređuje pristup vizuelnim podacima, ovaj sistem takođe predstavlja praktičan primer primene savremenih tehnika mašinskog učenja u realnim okruženjima, čime doprinosi razvoju naprednih rešenja u oblasti obrade i pretrage slika.

U zavisnosti od potreba korisnika, sistem podržava dva modela pretrage. Prvi model omogućava pretragu na osnovu slike, gde se kao ulazni parametar koristi fotografija, za koju će se u rezultatu vratiti deset najbližijih iz baze. Drugi model omogućava pretragu na osnovu tekstualnog upita, tako što se kao rezultat vraćaju fotografije koje vizuelno reprezentuju sadržaj zadanog teksta.

Sistem je realizovan kao veb aplikacija, pri čemu je pozadinski deo (*backend*) implementiran u *Python FastApi* radnom okviru, dok je korisnički interfejs implementiran u *React-u*. Ovakva arhitektura omogućava upotrebu savremene platforme za pretragu slika na osnovu opisa ili sličnih vizuelnih karakteristika, čime se značajno unapređuje korisničko iskustvo.

Aplikacija je osmišljena da bude intuitivna kako profesionalcima koji se bave analizom vizuelnog sadržaja, tako i krajnjim korisnicima koji žele da lakše i brže dođu do relevantnih vizuelnih informacija.

1.1. Skup podataka

U ovom radu korišćen je *Flickr30k Entities* [3] skup podataka, proširena verzija popularnog *Flickr30k* skupa, standardnog benchmarka za opisivanje slika. Skup obuhvata 31.800 fotografija i 158.915 rečeničnih opisa, organizovanih u fajlu *results.csv* sa kolonama: *image_name*, *comment_number* i *comment*. Nakon eksplorativne analize podataka zaključene su karakteristike skupa podataka. Glavni nalazi uključuju ravnomernu distribuciju komentara po slikama, umeren sentiment i kraće opise u većini komentara. Skup podataka je sam po sebi dobro pripremljen, tako da faza preprocesiranja ima svega nekoliko dodatnih koraka.

1.2. Radni tok vektorske baze podataka i CLIP modela mašinskog učenja

CLIP model mašinskog učenja predstavlja savremeni multimodalni pristup zasnovan na *Transformer* arhitekturi, sposoban za istovremenu obradu tekstualnih i vizuelnih podataka. Obuka *CLIP* modela obuhvata velike skupove tekstualno-vizuelnih parova, gde model uči zajednički reprezentativni prostor. Ovaj prostor omogućava da slični tekstualni i vizuelni sadržaji budu bliski u vektorskom prostoru, što čini *CLIP* izuzetno korisnim za zadatke kao što su pretraga, klasifikacija i semantičko povezivanje.

Kada se *CLIP* model integriše sa vektorskom bazom podataka, dobija se optimalan radni tok koji kombinuje snage oba sistema. Na početku, *CLIP* model se prethodno obučava kako bi naučio vektorske reprezentacije koje mogu da povežu tekst i slike na semantičkom nivou. Ove vektorske reprezentacije zatim se ugrađuju u vektorsku bazu podataka, koja omogućava efikasno skladištenje i brzo pretraživanje.

Kada korisnik unese tekstualni upit ili dostavi sliku, *CLIP* generiše vektorsku reprezentaciju upita. Vektorska baza zatim koristi napredne algoritme za približnu pretragu najbližih suseda, kao što je *HNSW*, kako bi pronašla slične vektore iz skupa podataka. Rezultati pretrage se vraćaju korisniku, pri čemu *CLIP* osigurava da ovi rezultati odgovaraju semantičkom kontekstu korisničkog upita.

Integracija *CLIP* modela i vektorske baze podataka omogućava značajno poboljšanje u odnosu na upotrebu samo jedne od ovih komponenti. Ako se koristi samo *CLIP*, rezultati mogu biti manje tačni u radu sa velikim skupovima podataka. Sa druge strane, ako se koristi samo vektorska baza podataka, rezultati mogu biti tehnički precizni, ali ne i u potpunosti semantički relevantni za korisnika. Kombinacija ove dve tehnologije omogućava postizanje i preciznosti i prilagođenosti rezultata korisničkim očekivanjima, čime se postiže optimalan balans između efikasnosti i korisničkog iskustva u multimodalnim aplikacijama.

1.3. Kontrastivno pretreniranje

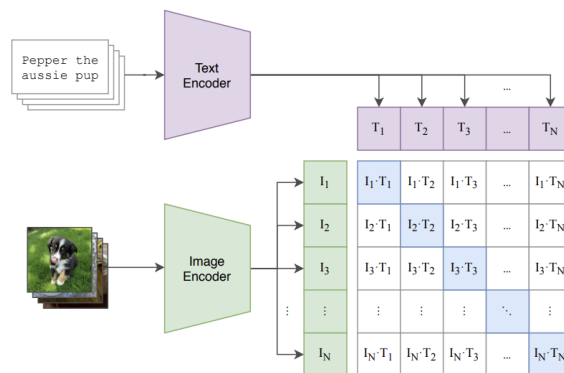
Kontrastivno učenje je metod obuke *AI* modela koji ga uči da razlikuje slične od različitih stvari, koristeći kontrastivni gubitak. Konkretnije, na osnovu skupa od N parova (slika, tekst), *CLIP* zajednički trenira enkoder za slike i enkoder za tekst kako bi predvideo koji od $N \times N$ mogućih (slika, tekst) parova u okviru skupa zapravo postoje [4].

Model prima *batch* podataka u obliku parova (slika, tekst). Za svaku sliku u *batch*-u, vizuelni enkoder generiše vektor slike. Prva slika odgovara vektoru I_1 , druga I_2 , itd. Svaki vektor je veličine de , gde je de dimenzija latentnog prostora, koja predstavlja broj komponenti (dimenzija) vektora koji opisuju sliku ili tekst u zajedničkom embedding prostoru. Veličina de je ključna jer određuje kapacitet modela da uči i reprezentuje kompleksne odnose između tekstualnih i slikovnih podataka. Veće dimenzije omogućavaju bogatiju reprezentaciju, ali istovremeno zahtevaju više resursa za obradu i memoriju. Izlaz ovog koraka je matrica $N \times de$. Slično, tekstualni opisi se konvertuju u tekstualne embedinge $\{T_1, T_2, T_3 \dots T_n\}$, čime se dobija matrica $N \times de$. Nakon toga, ove matrice se međusobno množe, izračunavajući kosinusne sličnosti između svake slike i teksta, što rezultira $N \times N$ matricom.

CLIP uči multimodalni ugnježdeni prostor tako što zajednički trenira enkoder za slike i enkoder za tekst kako bi maksimizirao kosinusnu sličnost ugnježđenja slike i teksta za N stvarnih parova u skupu, istovremeno minimizujući kosinusnu sličnost ugnježđenja za $N \times N - N$ pogrešnih parova.

Na slici 1.3.1, to znači da želimo da maksimiziramo dijagonalu u matrici, dok minimiziramo sve ostale elemente.

(1) Contrastive pre-training



Slika 1. Kontrastivno pretreniranje

Za svako ulazno slikovno ugnježđenje vrši se klasifikacioni zadatak u pravcu x ose, koji nam govori koji od tekstualnih ugnježđenja $\{T_1, T_2, T_3 \dots T_n\}$ najviše odgovara toj slici. Takođe, za svako tekstualno ugnježđenje vrši se klasifikacioni zadatak u pravcu y ose koji govori koje od slikovnih ugnježđenja $\{I_1, I_2, I_3 \dots I_n\}$ ga najbolje reprezentuje. Na primer, slika I_1 se opisuje tekstem T_1 , a ne tekstovima T_2, T_3 itd.

Gubitak koji se ovde koristi je *cross entropy* gubitak, izračunat između logita i labela u x pravcu i y pravcu, a zatim se računa njihova prosečna vrednost. To znači da ova vrsta gubitka minimizuje greške u oba pravca – od slike ka tekstu i od teksta ka slici. Suštinski, sa kontrastivnom tehnikom, *CLIP* je treniran da razume da slična predstavljanja treba da budu blizu u latentnom prostoru, dok različita treba da budu udaljena.

1.4. Pinecone

Pinecone [2] je specijalizovana baza podataka za upravljanje vektorima, dizajnirana da olakša rad sa velikim skupovima podataka i složenim zadacima pretrage u domenu veštačke inteligencije i mašinskog učenja. Zasnovana je na konceptu vektorske pretrage i omogućava korisnicima da efikasno organizuju, indeksiraju i pretražuju podatke visoke dimenzionalnosti.

Jedna od ključnih prednosti *Pinecone*-a je njegova sposobnost da precizno i brzo vrši pretragu sličnosti u velikim skupovima podataka. Ovo se postiže korišćenjem sofisticiranih algoritama za vektorsku pretragu, koji omogućavaju brzo pronalaženje najrelevantnijih rezultata na osnovu sličnosti ugrađenih vektora.

1.5. Hijerarhijski navigabilan mali svet (*HNSW*)

Hijerarhijski navigabilan mali svet [5] je savremena tehnika za pronalaženje približnih najbližih suseda datog vektora u velikom skupu vektora. Ona funkcioniše tako što gradi graf koji povezuje vektore na osnovu njihove

sličnosti ili udaljenosti, a zatim koristi strategiju pretrage zasnovanu na pohlepnom pristupu za kretanje kroz graf i pronalaženje najslbližijih vektora.

Algoritam takođe gradi hijerarhijsku strukturu grafa, gde se svakoj tački dodeljuje različit sloj sa određenom verovatnoćom. Viši slojevi sadrže manji broj tačaka i duže ivice, dok niži slojevi sadrže veći broj tačaka i kraće ivice. Najviši sloj sadrži samo jednu tačku, koja predstavlja ulaznu tačku za pretragu.

Hijerarhijska struktura omogućava efikasnu i preciznu pretragu, počevši od najvišeg sloja i postepeno prelazeći na niže slojeve, pri čemu se u svakom koraku bira najbliži čvor kao sledeća tačka pretrage.

Hijerarhijska struktura koju koristi HNSW omogućava brzu i preciznu pretragu, jer se pretraga započinje od najvišeg sloja i postepeno se spušta ka nižim slojevima, pri čemu se u svakom koraku bira najbliži čvor kao sledeća tačka pretrage.

Performanse HNSW algoritma zavise od nekoliko faktora, kao što su dimenzionalnost vektora, broj slojeva, broj suseda po čvoru i broj koraka pretrage. Ovi faktori utiču na balans između tačnosti i efikasnosti, kao i na složenost i skalabilnost algoritma.

2. SPECIFIKACIJA SISTEMA ZA PRETRAGU SLIKA

Sistem za pretragu slika koristi arhitekturu koja objedinjuje više komponenti i tehnologija za omogućavanje efikasnog indeksiranja i pretrage. Kao što je prikazano na dijagramu sa slike 2, ova arhitektura je zasnovana na integraciji klijentske aplikacije, CLIP modela i vektorske baze podataka sa sistemom za pretragu.



Slika 2. Arhitektura sistema

Klijentska aplikacija: Razvijena u React-u, ova komponenta pruža korisnički interfejs koji omogućava interaktivnu pretragu i prikaz rezultata. Korisnik može pretraživati slike na osnovu teksta ili druge slike, a aplikacija šalje zahteve sistemu za pretragu.

Sistem za pretragu: Glavni deo sistema, implementiran u Python-u, vrši osnovne funkcije za pretragu i obradu podataka. On komunicira sa CLIP modelom i vektorskom bazom kako bi kreirao i upravljao embedinzima za slike i tekst.

CLIP model: Ovaj model, razvijen od strane OpenAI-a, služi za kreiranje multimodalnih embeddinga koji predstavljaju semantički sadržaj slika i tekstova. CLIP model generiše embedinge koji se kasnije upoređuju sa embedinzima u vektorskoj bazi kako bi se našle slike koje su najslbližije zadanom upitu.

Vektorska baza (Pinecone): Pinecone služi kao baza ugnježđenja, gde se embedinzi slika i tekstova čuvaju u vektorskom formatu radi efikasnog pretraživanja. Kada

korisnik postavi upit, sistem upoređuje embedinge upita sa embedinzima u Pinecone bazi kako bi vratio najslbližije srezultate.

Ovakva arhitektura omogućava fleksibilno pretraživanje slika, bilo na osnovu tekstualnih upita ili sličnih slika, sa velikom tačnošću i brzinom.

3. EKSPERIMENT

Ovim poglavljem definisaćemo izabran način za evaluaciju implementiranog sistema i objasniti razlog za izbor istog.

U sistemima za pretragu informacija, kao što su pretrage teksta, slika ili multimodalnih podataka, metrike poput top-N tačnosti igraju ključnu ulogu u proceni kvaliteta sistema. Top-N tačnost predstavlja procentualni udeo upita gde je relevantni rezultat (tj. dokument, slika ili drugi objekat) pronađen među prvih N rezultata rangiranih od strane sistema.

U informacionim sistemima, korisnici očekuju da relevantni rezultati budu rangirani među prvima. Visoka top-1 tačnost pokazuje sposobnost sistema da konzistentno vraća najrelevantniji rezultat kao prvi, dok visoke vrednosti top-3, top-5 ili top-10 tačnosti ukazuju na to da sistem efikasno identifikuje većinu relevantnih rezultata, iako oni možda nisu uvek rangirani na samom vrhu.

Zbog velikih količina podataka u našem sistemu (oko 31.800 slika), korisnička evaluacija bila je nepraktična, jer bi ispitanici morali da procenjuju rezultate bez uvida u ceo skup slika. Zbog toga smo se odlučili za automatsko testiranje metrikom top-N tačnosti, koja omogućava objektivnu procenu na reprezentativnom uzorku od 500 nasumično izabranih slika.

4. REZULTATI I TUMAČENJA

Dobijeni rezultati prikazuju procenat tačnosti sličnosti slika i tekstova. Vrednosti po sve četiri metrike date su narednom tabelom

Tabela 1. Rezultati tačnosti CLIP modela u predviđanju sličnosti slika i tekstova

Metrika tačnosti	Procenat tačnosti
Top-1 tačnost	61%
Top-3 tačnost	75%
Top-5 tačnost	85%
Top-10 tačnost	94%

Ovi rezultati ukazuju da model često greši u identifikaciji najboljeg podudaranja, ali sa većim N postiže znatno bolju tačnost. Niska top-1 tačnost od 61% ukazuje na ograničenja modela u direktnoj tekstualnoj pretrazi. Ipak, visoka tačnost u top-10 (94%) čini ga veoma pogodnim za inicijalnu fazu pretrage, gde se potencijalno relevantni rezultati mogu rangirati za dalju obradu.

CLIP model pokazuje veliki potencijal kao osnova za sistem pretrage, uz mogućnost poboljšanja kroz dodatno podešavanje parametara, izborom drugih arhitektura enkodera ili integracijom sa specifičnim algoritmima za fino rangiranje.

5. ZAKLJUČAK

U ovom radu predstavljen je sistem za pretragu slika zasnovan na CLIP modelu i vektorskim bazama podataka, koji omogućava efikasnu pretragu vizuelnog sadržaja na osnovu tekstualnih ili slikovnih upita. Razmatrajući izazove savremenih sistema za pretragu, kao što su ograničenja tekstualnih oznaka i potreba za skalabilnošću, implementiran je pristup koji kombinuje multimodalne embedinge i vektorske baze kako bi se obezbedila tačnost i brzina u obradi velikog broja podataka.

Sistem je omogućio dve vrste pretraga – na osnovu teksta i na osnovu slike – pri čemu je postignuta visoka tačnost u identifikaciji relevantnih rezultata, naročito u top-10 metrici (94%). Iako top-1 tačnost od 61% ukazuje na potrebu za dodatnim poboljšanjima, rezultati su u skladu sa očekivanjima za primenu modela u složenim i velikim skupovima podataka.

Naučni doprinos rada ogleda se u uspešnoj primeni *CLIP* modela za rešavanje stvarnih problema pretrage vizuelnog sadržaja. Evaluacija sistema pruža vredne uvide u mogućnosti i ograničenja trenutnog pristupa, dok sposobnost modela da generalizuje preko različitih zadataka bez dodatnog finog podešavanja potvrđuje njegovu robusnost i svestranost.

5.1. Dalji razvoj sistema

Buduća istraživanja mogu se usmeriti na unapređenje tačnosti kroz primenu naprednih tehnika, uključujući različite arhitekture enkodera i pažljiv odabir parametara modela u skladu sa specifičnim zahtevima domena primene. Proširenje sistema detaljnijim pretprocesiranjem slika takođe bi moglo doprineti poboljšanju performansi.

Trenutno implementirani sistem čuva slike sa metapodacima u indeks vektorske baze, pri čemu metapodaci obuhvataju isključivo nazive slika. Unapređenje sistema dodavanjem opisa slike kao dela metapodataka značajno bi poboljšalo korisnost i funkcionalnost. Ovo proširenje bi, pored bolje pretrage, omogućilo i primenu naprednih tehnika semantičkog i hibridnog pretraživanja koje podržava *Pinecone*, čime bi se dodatno unapredile mogućnosti povezivanja i identifikacije relevantnih sadržaja.

Dodatno, integracija *reranker*-a u postojeći sistem mogla bi značajno poboljšati redosled prikazanih rezultata, naročito u slučajevima gde je potrebno preciznije rangiranje u okviru top rezultata. *Reranker* bi omogućio finu obradu i prilagođavanje rezultata specifičnim kontekstima pretrage, što bi ujedno podiglo nivo tačnosti sistema i korisničko iskustvo.

Zaključno, sistem razvijen u ovom radu predstavlja značajan korak napred u oblasti pretrage vizuelnog sadržaja, objedinjujući savremene tehnike mašinskog učenja i vektorske baze podataka. Njegova primena i rezultati pokazuju potencijal za široku upotrebu u različitim domenima, dok istovremeno postavljaju osnovu za budući razvoj i istraživanje u ovoj oblasti.

6. LITERATURA

- [1] openai.com, „<https://openai.com/index/clip/>“ [Na mreži].
- [2] pinecone.io, „<https://www.pinecone.io/>“ [Na mreži].
- [3] J. R, „kaggle.com,“ [Na mreži]. Available: <https://www.kaggle.com/datasets/hsankesara/flickr-image-dataset/data>
- [4] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.
- [5] Y. A. Malkov and D. A. Yashunin, "Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs," IEEE transactions on pattern analysis and machine intelligence, vol. 42, no. 4, p. 824–836, 2018.

Kratka biografija:



Katarina Artukov rođena je 12.06.2000. godine u Sremskoj Mitrovici. Školske 2019/20 godine upisuje se na Fakultet tehničkih nauka na studijski program Računarstvo i automatika, a 2022. godine upisuje usmerenje Primenjene računarske nauke i informatika. Osnovne akademske studije završila je u septembru 2023. godine, nakon čega je u istoj školskoj godini upisala master akademske studije.