Vizualno prepoznavanje lokacija (VPR - Visual place recognition)

1st Karmela Arambašić
Fakultet elektrotehnike i računarstva
Zagreb, Croatia
karmela.arambasic@fer.hr

2nd Kristo Palić
Fakultet elektrotehnike i računarstva
Zagreb, Croatia
kristo.palic@fer.hr

3rd Nikolina Špehar
Fakultet elektrotehnike i računarstva
Zagreb, Croatia
nikolina.spehar@fer.hr

4th Ana Žanko
Fakultet elektrotehnike i računarstva
Zagreb, Croatia
ana.zanko@fer.hr

5th Veronika Žunar
Fakultet elektrotehnike i računarstva
Zagreb, Croatia
veronika.zunar@fer.hr

***Sažetak*—Cilj ovog projekta je istražiti i usporediti različite modele za vizualno prepoznavanje lokacija (engl. Visual Place Recognition, VPR). U ovom radu analizirat će se napredak u dubokom učenju i računalnom vidu koji je značajno unaprijedio sposobnosti VPR sustava. Istraživanjem povezanih radova i analizom performansi različitih modela u radu bit će istaknuti zaključci o najefikasnijim metodama primijenjivih za navedeni problem.**

***Index Terms*—VPR, prepoznavanje mjesta, CNN, ekstrakcija značajki, GSV-Cities**

I. Uvod

Primjena VPR tehnologije proteže se kroz različite domene, uključujući autonomna vozila, robotsku navigaciju, geolokaciju u urbanim područjima i proširenu stvarnost. Ova tehnologija omogućava sustavima da prepoznaju i lociraju specifična mjesta u fizičkom okruženju koristeći vizualne informacije, što je od velikog značaja za navigaciju i interakciju s okolinom. Napredak u području dubokog učenja i računalnog vida značajno je unaprijedio sposobnosti VPR sustava. Modeli konvolucijskih neuronskih mreža (engl. Convolutional Neural Network, CNN), kao što su ResNet, DenseNet, VGG i specijalizirani modeli poput NetVLAD i PatchNetVLAD, omogućavaju ekstrakciju i agregaciju značajki koje su ključne za prepoznavanje mjesta.

II. Pregled metoda za VPR

U ovom dijelu izvještaja pružit ćemo pregled nekoliko metoda koje se često koriste u VPR-u, uključujući ResNet, DenseNet, NetVLAD i TransVPR.

A. ResNet

ResNet (Residual Network) je duboka konvolucijska neuronska mreža koja je poznata po uvođenju "rezidualnih" veza, što omogućuje izgradnju vrlo dubokih mreža bez problema s degradacijom performansi. ResNet se pokazao vrlo učinkovitim u raznim zadacima prepoznavanja slika, uključujući VPR. Osnovna ideja je da se izlaz svake složenije funkcije direktno prenosi u sljedeći sloj, zajedno s neizmijenjenim ulazom, što omogućuje učinkovitije učenje i bolje prepoznavanje lokacija na temelju vizualnih informacija.

B. DenseNet

DenseNet (Densely Connected Convolutional Network) dodatno poboljšava propagaciju informacija i gradijenata u mreži spajanjem svakog sloja sa svim prethodnim slojevima. Ova gusto povezivana struktura omogućuje bolju iskorištenost značajki i učinkovitije učenje, što rezultira boljim performansama u zadacima prepoznavanja slika. Za VPR, DenseNet može pružiti bogatiji skup značajki koje su ključne za točno prepoznavanje i razlikovanje različitih lokacija.

C. NetVLAD

NetVLAD je metoda koja kombinira prednosti tradicionalnog VLAD (Vector of Locally Aggregated Descriptors) deskriptora s dubokim učenjem. U NetVLAD-u, VLAD grupiranje je integrirano kao sloj unutar konvolucijske neuronske mreže, omogućujući krajnje do krajnje treniranje na ciljnim zadacima. NetVLAD je posebno koristan za VPR jer omogućuje učinkovito agregiranje lokalnih deskriptora u snažnu globalnu reprezentaciju slike, što olakšava prepoznavanje lokacija čak i u promjenjivim uvjetima.

D. TransVPR

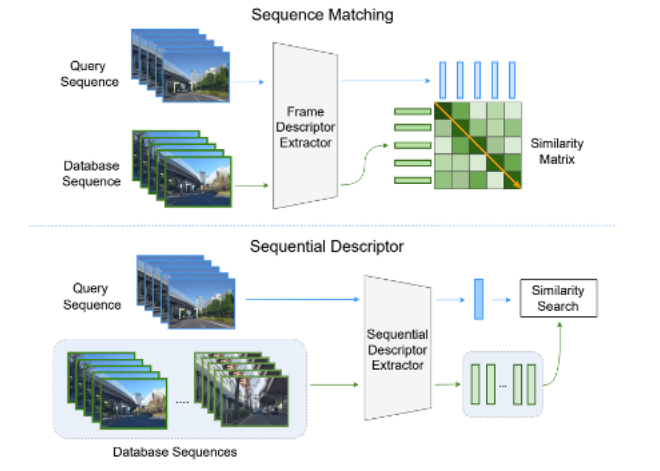
TransVPR je metoda koja koristi arhitekturu transformera prilagođenu za vizualno prepoznavanje lokacija. Za VPR, TransVPR može analizirati sekvence slika ili videozapisa, omogućujući modelu da uzme u obzir prostorne i vremenske informacije. Ovo je posebno korisno za prepoznavanje lokacija u dinamičnim ili složenim okruženjima gdje su promjene u vizualnim informacijama učestale. TransVPR može pružiti robusnije i preciznije prepoznavanje lokacija zahvaljujući svojoj sposobnosti da integrira i analizira bogate vizualne podatke.

III. Povezani radovi

Mana korištenja konvolucijskih mreža je ta da se tijekom uzorkovanja slojeva, kako bi se dobile informacije korisne za prepoznavanje lokacije, gube detalji slike. Taj problem može se rješiti korištenjem transformera prilikom rješavanja VPR problema.

Prema istraživanju [2] koje je razvijalo vlastiti model koji se za VPR nije oslanjao samo na konvolucijske mreže, nego ih je pokušao usporediti i poboljšati uporabom transformera, transformeri su se pokazali puno bolji u prepoznavanju različitih uvjeta na slici. Njima se mogu povezati dvije slike kao jednake iako je ambijent drugačiji, npr. jedna je slikana po danu, a druga po noći. Transformeri koriste mehanizam pažnje za obradu svih dijelova ulaznog podatka odjednom, omogućavajući direktne veze između udaljenih značajki. Korištenjem mehanizama uparivanja bitnih dijelova slike, transformeri imaju mogućnost povezivanja I uspoređivanja između svih slika lokacije. "Vision" transformeri pronalaze glavna područja slika koja su idealna za dugoročno prepoznavanje lokacije, odnosno područja na koja promjena uvjeta minimalno utječe na raspoznavanje lokacije slike (kao npr. promjena godišnjeg doba ili drugo doba dana).

Podudaranje sekvenci (engl. sequence matching) započinje tako da se svaki dio ulazne slike individualno usporedi s kolekcijom slika poznatih lokacija kako bi se izgradila matrica sličnosti. Izgradnja matrice sličnosti se u velikoj mjeri oslanja na pretpostavku jednakih uvjeta na slikama (npr. osvijetljenje, vremenski uvjeti, godišnje doba, itd.) što uvodi veliku mogućnost pojave grešaka. Također, korištenje ovog postupka neisplativo je zbog cijene obavljanja operacija koja linearno raste u ovisnosti o veličini mape i o duljini sekvence. Alternativna metoda koja se koristi umjesto metode podudaranja sekvenci oslanja se na deskriptore sekvenci (engl. sequential descriptors) koji imaju mogućnost sažimanja cijele sekvence. Time je omogućena brza pretraga sličnosti. Usporedba ove dvije metode vidljiva je na slici 1.



Slika 1. Podudaranje sekvenci i deskriptori sekvenci

Generiranje globalnih deskriptora važan je korak tijekom rješavanja VPR problema. Starije metode za generiranje oslanjale su se na agregaciju lokalnih deskriptora korištenjem tehnika kao što su "Vreća riječi" (engl. Bag of Words, BoW) i "Vektor lokalno agregiranih deskriptora" (engl. Vector of Locally Aggregated Descriptors, VLAD). Uz razvoj dubokog učenja, razvile su se metode koje povezuju stare tehnike s konvolucijskim mrežama, kao što je NetVLAD. Korištenjem konvolucijskih mreža za generiranje globalnih deskriptora, mreža sažima uzorke te se zbog toga gube bitni detalji iz ulaznih podataka. Iz tog razloga su se za generiranje globalnih transformera počeli koristiti transformeri. Transformeri se koriste na način da se skupljalju tokeni manjih površina slike te se spajaju u sveobuhvatniju reprezentaciju cijele ulazne slike [2].

IV. Materijali i metode

A. Dataset

GSV-Cities je opsežan dataset dizajniran za istraživanja i aplikacije u području VPR-a. Sastoji se od velikog broja slika urbanih okruženja. Dataset je kreiran s ciljem da podrži razvoj i evaluaciju algoritama za prepoznavanje mjesta, posebno u varijabilnim i dinamičnim urbanim okruženjima. Objavljen je u "Neurocomputing 2022: GSV-Cities: Toward Appropriate Supervised Visual Place Recognition".

Sadrži oko 530 tisuća slika s više od 62 tisuće različite lokacije, raspoređenih diljem različitih gradova svijeta. Svaka lokacija prikazana je s najmanje četiri, a najviše dvadeset slika. Sve su lokacije fizički udaljene (najmanje 100 metara između dvije lokacije).

Dataset je organiziran tako da svako mjesto ima dodijeljen ID, a slike koje prikazuju isto mjesto grupirane su zajedno pod istim ID-om. Prikazani podaci također uključuju orijentacije (engl. bearings) svake slike, što omogućuje prikazivanje istog mjesta iz istog ugla, ali u različitim vremenima.

Postoje tri glavna problema s kojima se obično suočavaju datasetovi sa slikama:

- Geografska pokrivenost - Većina postojećih datasetova sakupljena je u malim područjima, od grada do male četvrti, što ih čini nedovoljno velikima za obuku na velikoj skali.
- Preciznost referentnih podataka - Veći datasetovi često imaju netočne referentne podatake, što je problem za nadzirano učenje. GSV-Cities rješava ovaj problem osiguravanjem visoko preciznih GPS koordinata i smjera snimanja za svaku sliku, omogućujući stvaranje preciznih pozitivnih i negativnih parova za trening.
- Perceptualna raznolikost - Mnogi datasetovi ne pružaju dovoljno varijacija u izgledu, poput promjena u perspektivi ili strukturalnih promjena. GSV-Cities uključuje slike snimljene u različitim vremenskim uvjetima, različitim godišnjim dobima i pod različitim uvjetima osvjetljenja, čime se osigurava visoka razina perceptualne raznolikosti.

B. Vlastita implementacija Resnet-18 modela

Za potrebe implementacije vlastitog modela odabrali smo Resnet-18 model. ResNet-18 je duboka konvolucijska neuronska mreža koja se sastoji od 18 slojeva, raspoređenih u četiri grupe rezidualnih blokova. Glavne komponente modela su:

- Ulazni sloj:
  - Konvolucijski sloj
  - normalizacija podataka
  - ReLU aktivacijska funkcija
  - Max Pooling
- Rezidualni blokovi:
  - Prva grupa (64 izlaznih kanala)
  - Druga grupa (128 izlaznih kanala)
  - Treća grupa (256 izlaznih kanala)
  - Četvrta grupa (512 izlaznih kanala)
- Izlazni sloj:
  - Sloj globalnog sažimanja (engl. Global Average Pooling)
  - Potpuno povezani sloj (engl. Fully Connected)

Rezidualni blok u ResNet-18 sastoji se od:

- Dva 3x3 konvolucijska sloja
- normalizacije podataka nakon svakog konvolucijskog sloja
- ReLU aktivacija nakon normalizacije podataka
- Spojnica (engl. shortcut) koja dodaje ulaz bloka na izlaz bloka

Iako postoje različite inačice ResNet modela, odlučili smo se za najjednostavniji model zbog ograničenih računalnih resursa i iznimne kvalitete podataka. Postupak vlastite implementacije uključuje sljedeće korake:

- priprema podataka
- odabir optimalnih hiperparametara
- definiranje modela
- treniniranje modela
- evaluacija modela

1) Priprema podataka: Prvo smo definirali prilagođeni dataset `CustomImageDataset` za učitavanje slika iz direktorija. Na slike smo zatim primijenili sljedeće transformacije: promjena veličine na 224x224 piksela, nasumično horizontalno okretanje, normalizacija te konverzija u tenzore.

2) Odabir optimalnih hiperparametara: Zbog ograničenih računalnih resursa nismo imali mogućnost eksperimentiranja hiperparametrima do razine koje smo htjeli, ali smo uspjeli testirati modele sa različitim stopama učenja, različitim optimizacijskim algoritmima (SGD i Adam) te različitim veličinama serija (engl. batch). Kombinacije modela testirali smo na 5% podataka kako bi dobili uvid koji parametri su najbolji za naš model.

3) Definicija modela: Za definiranje modela koristili smo unaprijed trenirani ResNet18 model s prilagođenim završnim slojem za broj klasa u našem datasetu. Također, implementirali smo optimalne parametre iz prethodnog koraka.

4) Trening modela: Model smo trenirali pomoću funkcije `trainModel` koja obuhvaća unakrsnu validaciju i `EarlyStopping`. Funkcija `trainModel` iterira kroz unaprijed definirani broj epoha (50), trenira model na trening skupu podataka, validira ga na validacijskom skupu te koristi `EarlyStopping` kako bi prekinula trening ukoliko nema poboljšanja u gubitku na validacijskom skupu kroz određeni broj epoha.

5) Evaluacija modela: Model smo evaluirali na trening, validacijskom i testnom skupu pomoću funkcije `evaluateModel`. Funkcija `evaluateModel` procjenjuje model na danom skupu podataka te vraća ukupni gubitak i točnost.

V. Rezultati

Nakon izvršenog treniranja, naš model je postigao sljedeće rezultate:

- Trening gubitak: 0.0308, točnost: 99.20%
- Validacijski gubitak: 0.2678, točnost: 91.49%
- Test gubitak: 0.2693, točnost: 91.53%

Rezultati pokazuju visoku točnost modela na trening, validacijskom i test skupu, što ukazuje na dobre generalizacijske sposobnosti modela.

VI. Zaključak

Razvoj u području dubokog učenja, uz napredak konvolucijskih neuronskih mreža i transformera, značajno je pridonio poboljšanju mogućnosti VPR sustava. U ovom radu prestavljen je sveobuhvatni pregled VPR-a kroz objašnjenje učestalih metoda i njihovih primjena.
Našom implementacijom ResNet-18 modela pokazana je visoka učinkovitost konvolucijskog modela za problem VPR-a na raznolikom i opširnom GSV-Cities datasetu. S obzirom na ograničene resurse nismo uspjeli isprobati transformer arhitekturu, no daljnim istraživanjima i usavršavanjem modela temeljenih na takvoj arhitekturi moglo bi se postići još bolji rezultati i na kompleksnijm problemima. Široki spektar moguće primjene VPR-a čini ovu tehnologiju ključnom u modernom društvu te doprinosi mogućem poboljšanju svakodnevnog života.

Literatura

[1] R. Mereu, G. Trivigno, G. Berton, C. Masone, B. Caputo, "Learning Sequential Descriptors for Sequence-based Visual Place Recognition", 2022.

[2] S. Sundar Kannan, B. Min, "PlaceFormer: Transformer-based Visual Place Recognition using Multi-Scale Patch Selection and Fusion", 2024.

[3] A. Ali-bey, B. Chaib-draa, P. Giguère "GSV-CITIES: Toward Appropriate Supervised Visual Place Recognition"