Data-efficient representation learning for visual place recognition

Leyva Vallina, María

DOI:
10.33612/diss.736449452

Samenvatting

Dit proefschrift onderzoekt het probleem van visuele plaatsherkenning, dat een fundamenteel onderdeel is van veel visuele lokalisatiesystemen en daarom van grote waarde is voor de computer vision gemeenschap. In het bijzonder behandelen we twee problemen in het veld: de eerste helft van dit proefschrift is gewijd aan de presentatie en evaluatie van een nieuwe dataset in een zeer onderverkende soort omgeving, namelijk tuinomgevingen. De tweede helft van dit proefschrift richt zich op het algoritmische deel van visuele plaatsherkenning en stelt een paradigmaverschuiving voor om visuele descriptoren te leren die sterkere, betrouwbare en kwantificeerbare representaties van de similariteit van beelden coderen. De bijdragen van dit proefschrift zijn als volgt.

- In Hoofdstuk 2 introduceren we de TB-Places dataset, die ongeveer 60k afbeeldingen bevat die zijn genomen in twee tuinomgevingen. Met deze dataset vullen we de kloof van het gebrek aan data voor visuele lokalisatie in dit soort omgevingen, die zeer specifieke uitdagingen bieden vanwege het perspectief, het weer en de belichtingsvariaties in de beelden, evenals de aanwezigheid van repetitieve, zeer vergelijkbare texturen die de detectie van discriminerende visuele signalen zeer uitdagend maken. Tevens evalueren we bestaande kant-en-klare methoden voor visuele plaatsherkenning en analyseren we de verkregen resultaten.
- We breiden de TB-Places dataset uit met 8k extra afbeeldingen, zoals gepresenteerd in Hoofdstuk 3, en vervolgens breiden we de benchmark evaluatie uit met meer methoden en modellen die specifiek zijn afgestemd op deze taak. De resultaten die we verkregen suggereert dat er veel ruimte was voor verbetering en dat de bestaande visuele plaatsherkenningsmethoden een aantal fundamentele tekortkomingen vertoonden.
- In de tweede helft van dit proefschrift gaan we in op een specifieke zwakte die we ontdekten tijdens ons werk in de eerste twee hoofdstukken: bestaande algoritmen

voor visuele plaatsherkenning leren representaties door een binaire grondwaarheid van visueel gelijke of ongelijke beeldparen te definiëren en vervolgens een verliesfunctie te optimaliseren die in wezen bedoeld is om te onderscheiden tussen twee goed gedefinieerde en scheidbare klassen. Descriptoren voor visuele plaatsherkenning worden echter verondersteld representaties te coderen die een adequate maat van beeldovereenkomst geven wanneer twee beeldrepresentaties aan een bepaalde scorefunctie (d.w.z. de Euclidische afstand) worden gekoppeld. Dit is geen binair classificatieprobleem, aangezien beeldovereenkomst niet discreet is, maar continu: twee beelden kunnen volledig vergelijkbaar of volledig verschillend zijn, maar ook alles daartussenin. Op basis van dit inzicht herdefiniëren we de binaire grondwaarheid van bestaande visuele plaatsherkenningsdatasets (namelijk MSLS, TB-Places en 7Scenes) om een continue maat van beeldovereenkomst te coderen, berekend op basis van geometrische informatie. Vervolgens gebruiken we deze nieuwe grondwaarheid om een nieuwe Generalized Contrastive Loss te optimaliseren, die niet alleen gedefinieerd is voor de extreme gevallen van de grondwaarheid (0 en 1), maar rekening houdt met elke waarde € [0,1]. We presenteren onze nieuwe methode en resultaten in Hoofdstuk 4, en we laten zien dat we met een eenvoudige aanpassing de leerpijplijn kunnen vereenvoudigen, de trainingstijd kunnen verkorten en beter kunnen presteren dan de state-of-the-art methoden door robuustere descriptoren te leren.

- De Generalized Contrastive Loss leidt weliswaar tot uitstekende resultaten, maar introduceert nog steeds een kunstmatige binarisering van het probleem, omdat de functie uit twee termen bestaat: de eerste duwt de descriptoren van gelijksoortige afbeeldingen samen en de tweede trekt die van ongelijksoortige afbeeldingparen uit elkaar. Dit kan leiden tot prestatieproblemen, die we ontdekten bij het trainen van transformator backbones. We herdefiniëren de visuele plaatsherkenning dus volledig en benaderen het als een regressieprobleem: we regresseren alleen een parameter voor de gelijkenis van afbeeldingen (gecodeerd met de Euclidische afstand tussen twee representaties) en matchen deze met onze geannoteerde gelijkenisgrondwaarheid. Zoals we in Hoofdstuk 5 uitvoerig presenteren, kunnen we met deze paradigmaverschuiving in korte tijd transformator-backbones trainen en betere resultaten behalen dan wanneer we de GCL-functie gebruiken. Bovendien tonen we aan dat methoden die zijn getraind met behulp van regressie concurrerende prestaties bereiken in veel minder trainingsiteraties dan hun equivalenten die zijn getraind om een GCL te optimaliseren, wat een zeer data-efficiënte aanpak aantont.