Learning Deep Representation for Place Recognition in SLAM

Made By:
Ajinkya Bedekar (U101116FCS183) (D5)
Devansh Anhal (U101116FCS031) (D2)
Dhruva Agarwal (U101116FCS177) (D5)
Harsha Deuri (U101116FCS043) (D2)

Understanding of the Paper

- Place Recognition
  - accurately and efficiently recognize the location of a given query image
- SLAM
  - computational problem of constructing or updating a map of an unknown environment while simultaneously keeping track of an agent's location within it
- Loop Closure Detection
  - reconstruct three-dimensional environments and estimate a camera trajectory accurately
- Pose Graph Optimization
  - estimate robot trajectory (collection of poses) from relative pose measurements
- Deconvolution Net
  - composed of deconvolution and unpooling layers
- KITTI Visual Odometry Dataset
- New College Dataset

Introduction

- SLAM is an important task in the context of robot navigation with vision
- Entire SLAM process relies on recognizing places the robot has already visited to achieve visual loop closure detection
- Major tasks:
  - representing the frames with the help of visual descriptors
  - judging the similarity between the frames based on the descriptors
- In the context of this work, place recognition refers to recognising whether a place has been visited previously or not.

Background and/or Related Work

BoW Based Approaches

- Fixed size vocabulary is used as a vector quantizer to classify descriptors in an image frame.
- FABMAP model considers a sequence of non-overlapping frames and checks if each frame belongs to an already visited place. It suffers from the problem of perceptual aliasing.
- Methods like SeqSLAM perform correlation-based matching on short sequences of images instead of depending directly on individual image frames.
- Voting based methods perform a nearest neighbour search on the image descriptor space to identify potential matches.
- Image descriptors like SIFT, BRISK or FREAK are also used to form the descriptor vector.

Background and/or Related Work

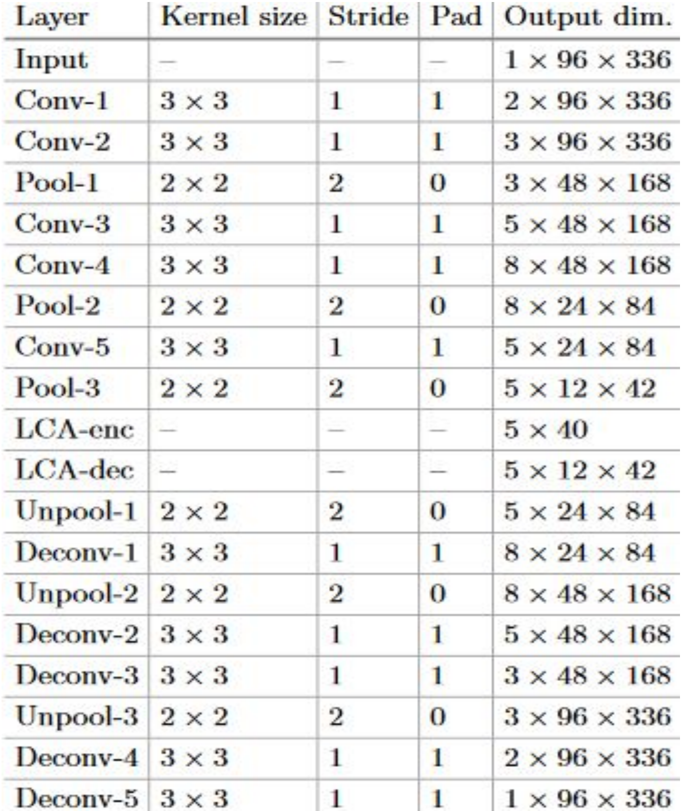
Deep Learning Based Approaches

- Convolutional neural network (CNN) based approaches have been developed for loop closure detection.
- Overfeat network trained on the ImageNet dataset is used to extract features from the image frames.
- In this approach the network was pre-trained on the ImageNet dataset and thus it is optimized mainly for object recognition and not oriented towards place recognition as desired for loop closure detection.
- Denoising autoencoders (DA) have also been used for localization tasks.

Proposed Model in Paper/Models Used in the Paper

- An autoencoder based deep learning network that extracts a lower dimensional vector representation of an image.
- With an autoencoder trained to encode and decode an image, the task of loop detection reduces to finding the distance between the encoded vectors of the query image and the input image.
- Whenever the distance falls below a certain threshold a loop closure can be reported.
- The value of the threshold can be either learned or tuned based on previous experience about the alteration limits of the environment.
- The reconstruction process in this case uses the concept of switch matrix which holds the position of the pixel selected during a pooling layer of the encoder so that proper mapping can be done during decoding.

Implementations of the Model



Layer	Kernel size	Stride	Pad	Output dim.
Input	–	–	–	1 × 96 × 336
Conv-1	3 × 3	1	1	2 × 96 × 336
Conv-2	3 × 3	1	1	3 × 96 × 336
Pool-1	2 × 2	2	0	3 × 48 × 168
Conv-3	3 × 3	1	1	5 × 48 × 168
Conv-4	3 × 3	1	1	8 × 48 × 168
Pool-2	2 × 2	2	0	8 × 24 × 84
Conv-5	3 × 3	1	1	5 × 24 × 84
Pool-3	2 × 2	2	0	5 × 12 × 42
LCA-enc	–	–	–	5 × 40
LCA-dec	–	–	–	5 × 12 × 42
Unpool-1	2 × 2	2	0	5 × 24 × 84
Deconv-1	3 × 3	1	1	8 × 24 × 84
Unpool-2	2 × 2	2	0	8 × 48 × 168
Deconv-2	3 × 3	1	1	5 × 48 × 168
Deconv-3	3 × 3	1	1	3 × 48 × 168
Unpool-3	2 × 2	2	0	3 × 96 × 336
Deconv-4	3 × 3	1	1	2 × 96 × 336
Deconv-5	3 × 3	1	1	1 × 96 × 336

Experiments (Detailing Out Your Implementations)

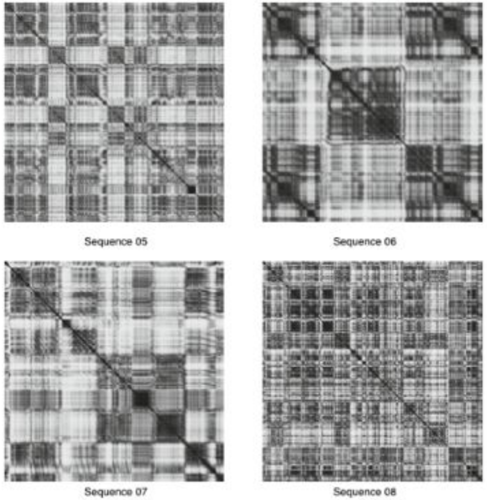


Fig. 1. Confusion matrices for sequences 5, 6, 7 and 8 of KITTI dataset. Darker the value, images are more similar.

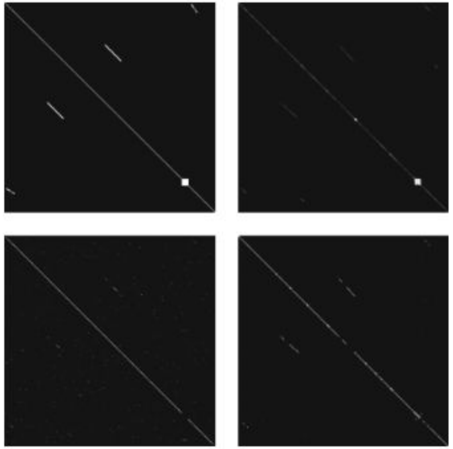
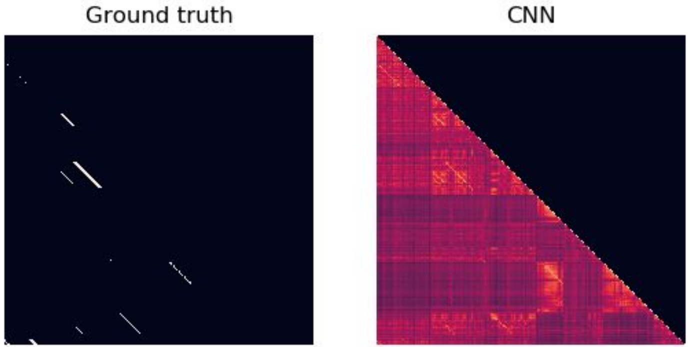
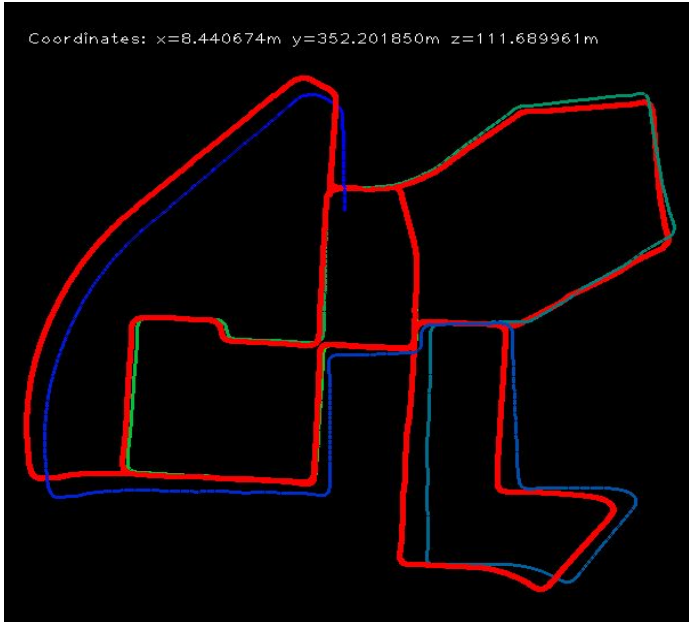


Fig. 2. Loop Closure Detection: ground truth matrix (top left) matrices for proposed methodology (top right), OpenSeqSlam [15] (bottom left) and OpenFabmap [4] (bottom right).

Experiments (Detailing Out Your Implementations)





```
Coordinates: x=8.440674m y=352.201850m z=111.689961m
```

Conclusion Derived from the Paper

- In this work the authors have proposed a deep learning autoencoder network that can represent an image with significantly lower dimension.
- In the approach discussed in paper, the authors tried to combine the best of both the deep learning approaches (weight sharing in CNNs and unsupervised feature learning in DAs) in a deconvolution net.
- The advantage of this approach is that vectors generated for two frames of the same scene which differ geometrically but are similar contextually and by content, are quite close to each other.
- Thus the approach works in general place recognition tasks also and can be extended to context and content based image matching problems.

Use-Case of the Work Done

- Autonomous Navigation Tasks
- Self Driving Car
- Unmanned Aerial Vehicles (UAV’s)
- Robot Vacuum Cleaners
- Gaming

Thank You!