# Image Pre-processing vs. Transfer Learning for Visual Route Navigation

William Smith, Yvan Petillot and Robert Fisher

Heriot-Watt University, University of Edinburgh

## Introduction

Neural networks (e.g. NetVLAD) are the state of the art for descriptor-based visual place recognition, but are affected by weather, lighting and man-made visual changes. NetVLAD is effective, but is computationally intensive. This work examines two techniques to address this problem.

Firstly, five image pre-processing techniques are used before the embedded descriptors are extracted using an ImageNet pre-trained network with no further training.

Secondly, inspired by NetVLAD, a novel 'FullDrop' layer is created and trained using novel triplet learning schemes for visual route navigation in a variety of conditions 3x faster with 30x fewer parameters than NetVLAD.

## Method

**Image Pre-processing**

Images from a single route in two different conditions were pre-processed using the six techniques (PREP 1-5) in Figure 1 and embedded image descriptors were extracted using VGG16 pre-trained on ImageNet data.

**Triplet Network**

Six front-facing, pre-recorded and geotagged example traversal videos of a single route are selected for training data. A custom 'FullDrop' layer is appended to VGG16 and MobileNet pre-trained on ImageNet data with the final classification layer removed. The FullDrop layer is trained to model the route with triplet learning. The choice of triplets for learning has a significant effect on performance. The FullDrop layer triplet mining differs from NetVLAD by using positive images from less than or equal to 10 adjacent frames from the anchor and negative images more than 10 adjacent frames away.

## Image Pre-processing

Image pre-processing techniques are applied before descriptor extraction: (1) adaptive histogram normalisation; (2) histogram normalisation; (3) histogram normalisation and subtraction of the image's mean/std. Deviation; (4) subtraction of the image's mean/std. deviation and (5) pixel-wise RGB normalisation.

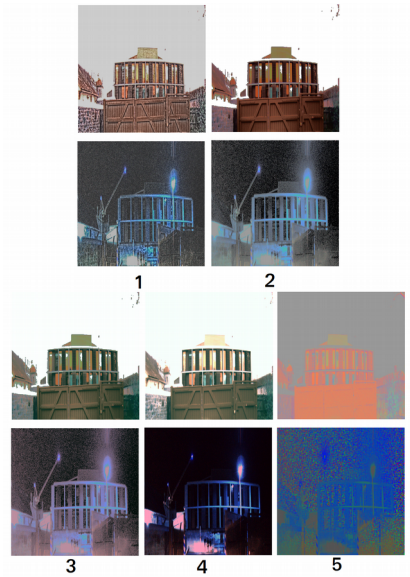


Figure 1: Pre-processing applied to night and day images.

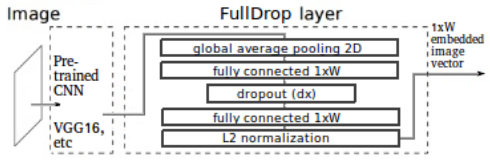


Figure 2: FullDrop layer architecture.

TABLE I

THE LOCALISATION ERROR (METRES) OF NIGHT AND DAY ROUTE TRAVERSALS, WITH INFERENCE TIME

| | $\mu$ (m) | Median (m) | $\sigma$ (m) | Inf.T. (ms) |
|---|---|---|---|---|
| MobileNet + FullDrop | 13.3 | 3.32 | 25.8 | 7 |
| NetVLAD | 15.5 | 2.39 | 29.1 | 80 |
| VGG16 + FullDrop | 49.4 | 9.57 | 73.8 | 10 |
| PREP 5 + VGG16 | 50.9 | 29.0 | 54.5 | 11 |
| PREP 1 + VGG16 | 70.1 | 54.4 | 66.4 | 10 |
| PREP 2 + VGG16 | 72.6 | 65.7 | 63.6 | 10 |
| PREP 4 + VGG16 | 73.0 | 56.4 | 65.2 | 11 |
| VGG16 | 83.6 | 78.5 | 62.1 | 10 |
| PREP 3 + VGG16 | 96.8 | 58.8 | 94.9 | 10 |

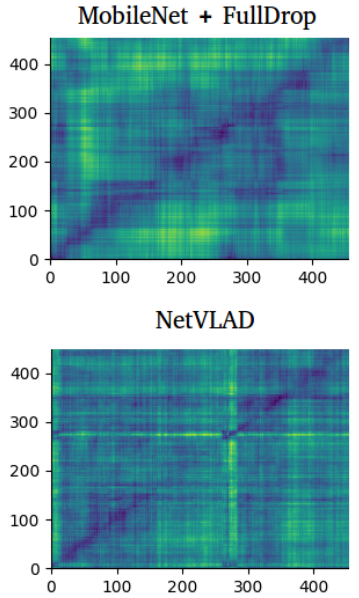


Figure 3: Partial difference matrices of night and day route descriptor comparisons from Table 1. Dark blue is low match error.

## Results

Night and day traversals of a 2km route from the Oxford RobotCar Dataset were used for evaluation in Table 1. Frames from each route were pre-processed and passed through the pre-trained VGG16 or just through the FullDrop or NetVLAD model to produce embedded image descriptors which were compared using the Euclidean distance; the 20 closest matches were generated and the one closest to the ground truth was identified as the localisation prediction.

PREP 5 reduces mean localisation error by 40.9%. The taught VGG16 + Full- Drop model reduces the median error by a further 67.0%. MobileNet + FullDrop reduces mean error by 14.2% compared to NetVLAD. The FullDrop model can generalise to unseen conditions and produces median localisation predictions that are within 7.2m of NetVLAD's.

Figure 3 shows FullDrop descriptors represent the similarities between similar sections of route more accurately than NetVLAD which shows a weaker relationship between adjacent frames. The results suggest FullDrop descriptors are better for utilising the relationship between consecutive frame for navigation, however this may introduce confusion between nearby, similar frames.

The MobileNet + FullDrop and NetVLAD models consisted of approximately 8x10^5 and 2.4x10^7 parameters respectively. Retraining NetVLAD takes at least 3x longer.

## Conclusion & Future Work

Reduced layer complexity decreases inference and training time while maintaining generalisation across a single route and resilience to the problem of perceptual aliasing. Image pre-processing improves performance and maintains decreased inference time with no need for re-training.

Future work aims to probabilistically filter the neural network descriptors to take advantage of the low median localisation error for accurate, real-time route navigation.

Acknowledgements

This research is supported by the Engineering and Physical Sciences Research Council (EPSRC), as part of the CDT in Robotics and Autonomous Systems at Heriot-Watt University and The University of Edinburgh.

Contact Details

Email: ws8@hw.ac.uk

Web: www.edinburgh-robotics.org