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

William H. B. Smith [1], Yvan Petillot [2] and Robert B. Fisher [3]

***Abstract*— This paper investigates image pre-processing and triplet learning for place recognition in route navigation. The first contribution combines image pre-processing and ImageNet pre-trained neural networks for generating improved image descriptors. The second contribution is a fast, compact 'FullDrop' layer that can be appended to an ImageNet pre-trained network and taught to generate invariant image descriptors with triplet learning. The proposals decrease inference time by 8x and parameters by 30x while keeping comparable performance to NetVLAD, the state of the art for this task**

I. INTRODUCTION

Neural network features provide state of the art recognition accuracy [1], but are affected by weather, lighting and man-made changes for visual place recognition [2]. This paper compares two techniques to create descriptors that are robust to these visual changes. Firstly, images are pre-processed before their descriptors are extracted by a pre-trained neural network. Secondly a 'FullDrop' layer is appended to the pre-trained network and trained with a custom triplet learning scheme to model a chosen route in a variety of conditions 3x faster with 30x fewer parameters than the state of the art for this task: NetVLAD [3]. Not all route appearances can be included in training data for deep learning solutions to this problem. The fast re-training possible with this paper's approach instead seeks to generalise across just one specified route in multiple conditions. The performance of the embedded descriptors generated by both approaches is compared with those generated by the pre-trained NetVLAD. The Oxford RobotCar Dataset [4] is used for evaluation. This paper's two contributions are:

1) Image pre-processing and neural network combinations for improved feature descriptors.
2) Triplet learning scheme and 'FullDrop' layer to generate image descriptors 8x faster with 30x fewer parameters than NetVLAD and comparable performance.

II. BACKGROUND

Initially, CNN's pre-trained on the ImageNet object recognition dataset were used to generate image descriptors for place recognition [5]. Novel architectures [6] were then trained end-to-end from scratch for place recognition. Triplet learning achieved positive results [7], but struggled to compete with off the shelf CNN descriptors.

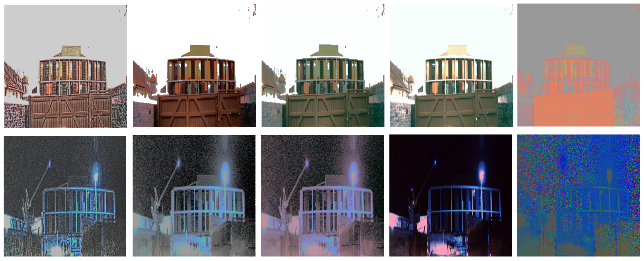

Fig. 1. Pre-processing methods (PREP 1-5) on day and night images, left to right: adaptive histogram normalisation; histogram normalisation; histogram normalisation and subtraction of the image's mean/std. deviation; subtraction of the image's mean/std. deviation; pixel-wise RGB normalisation.

NetVLAD [3] is the current state of the art in place image retrieval. VGG16 is frozen for training down to, but not including, the Conv5 layer then trained using triplet learning with a final VLAD layer. Performance is evaluated in a city environment at minimum intervals of 12 metres and in other environments at discrete locations. New pre-trained ImageNet networks, such as MobileNet have been released but can't be used in NetVLAD without lengthy re-training.

III. METHOD

A. Image Pre-processing

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

B. Triplet Network

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 [8]. 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.

Six video traversals of the same route: Early Evening, Morning, Midday, Rain, Afternoon and Overcast were sampled and synchronised for training. Each traversal was 884 frames at intervals of approximate 2.25m, making a total of 5304 training images. Approximately 51,000 triplets were mined for 10 training epochs in batches of 16 with a margin of 1. NetVLAD's training differed from the FullDrop training

This work was supported by the EPSRC Centre for Doctoral Training in Robotics and Autonomous Systems.

[1] William Smith, The University of Edinburgh and Heriot-Watt University, Edinburgh, EH14 4AS, UK whbsmith@gmail.com

[2] Yvan Petillot, Heriot-Watt University, Edinburgh, EH14 4AS, UK y.r.petillot@hw.ac.uk

[3] Robert B. Fisher, School of Informatics, University of Edinburgh, EH8 9LE, UK rbf@inf.ed.ac.uk

by unfreezing the Conv5 layer of the base network at training time and only used the hardest available triplets. Triplet learning relies on the raw difference between images so the FullDrop layer was not combined with image pre-processing.

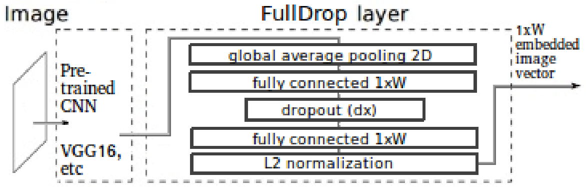

Fig. 2. Architecture of the system, including the 'FullDrop' layer. W is the width of the layer, which in this case was 512. The dropout parameter was 0.4 and L2 normalization was 0.1. This layer can be appended to any ImageNet pre-trained network and used for this task.

IV. Results

The NetVLAD model 'VGG-16 + NetVLAD + whitening' published online by [3] was used as a comparison. NetVLAD was pre-trained on the urban Pittsburgh 30k dataset and shown to be effective on a variety of different urban and suburban locations.

Night and day traversals of a 2km route synchronised and sampled to 878 images from the Oxford RobotCar Dataset taken months apart were used for the evaluation in Table IV-.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 in an effort to balance the potential accuracy of a probability-based system with pure place recognition.

1) Image Pre-processing: Four of the techniques reduced place recognition error. Specifically PREP 5 reduces mean error by 40.9% and median error by 63.1%.

TABLE I

THE LOCALISATION ERROR (METRES) OF NIGHT AND DAY ROUTE TRAVERSALS, WITH INFERENCE TIME AND TRAINING PARAMETER COUNT, AS DESCRIBED IN SECTIONS III-A AND IV

	μ (m)	Median (m)	σ (m)	Inf.T. (ms)	Params.
MobileNet + FullDrop	13.3	3.32	25.8	7	8x10^5
NetVLAD	15.5	2.39	29.1	80	2.4x10^7
VGG16 + FullDrop	49.4	9.57	73.8	10	5x10^5
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	-

2) FullDrop vs. NetVLAD: The taught VGG16 + FullDrop model reduces the median error from PREP 5 by a further 67.0%. MobileNet + FullDrop reduces mean error by 14.2% compared to NetVLAD. The FullDrop model is capable of generalisation to unseen conditions and produces median localisation predictions that are within 7.2m of NetVLAD's, which was extensively pre-trained for this task.

Figure 3 illustrates the comparison between the two evaluation routes' descriptors. The brief, but route specific training for the FullDrop descriptors allow them to represent the similarities between similar sections of route more accurately in comparison to NetVLAD which shows a weaker relationship between adjacent frames. The results suggest FullDrop descriptors are better for utilising the relationship between consecutive frame descriptors for navigation. FullDrop's more accurate model of the route may introduce confusion between nearby, similar frames.

The MobileNet + FullDrop and NetVLAD models consisted of approximately 8x10^5 and 2.4x10^7 parameters respectively. An Intel i7 CPU and Nvidia GTX 1070 trained the FullDrop layer in 510 seconds once the features had been extracted using the base network. Training a VLAD layer took 3x longer, re-training the full NetVLAD would take far longer.

3) Inference Time: NetVLAD (VGG16 + VLAD layer), VGG16 + FullDrop and MobileNet + FullDrop take 80ms, 10ms and 7ms to generate embedded descriptors on an Nvidia GTX 1070.

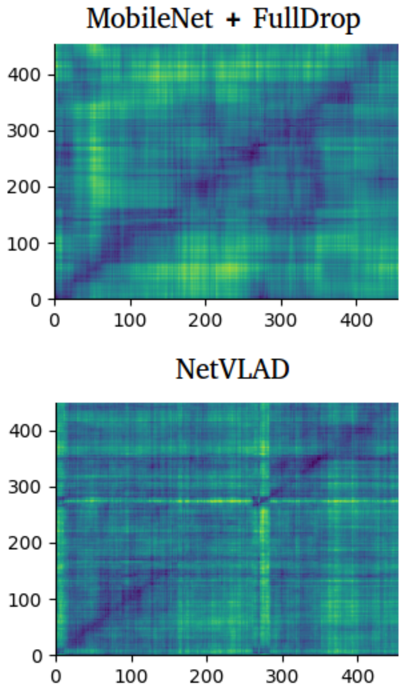

Fig. 3. Partial difference matrices of night and day route traversals from Table IV-.1. Showing the relative embedded descriptor relationships. The darker the blue the lower the match error.

V. Conclusion

Deep learning for discrete place recognition provides inspiration for real-time continuous route navigation. Reduced 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.

References

[1] A. Sharif, R. H. Azizpour, J. Sullivan, and S. Carlsson, “CNN Features off-the-shelf: an Astounding Baseline for Recognition,” in *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2014.

[2] S. Lowry, N. Sunderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, “Visual Place Recognition: A Survey,” *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 1–19, 2016.

[3] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, “NetVLAD: CNN Architecture for Weakly Supervised Place Recognition,” in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, 2018, pp. 1437–1451.

[4] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, “1 year, 1000 km: The Oxford RobotCar dataset,” *Int. J. of Robotics Research*, vol. 36, no. 1, pp. 3–15, 2017.

[5] T. Naseer, W. Burgard, and C. Stachniss, “Robust Visual SLAM Across Seasons,” in *IEEE Trans. on Robotics*, vol. 34(2):289-302, 2018.

[6] Z. Chen, A. Jacobson, N. Sunderhauf, B. Upcroft, L. Liu, C. Shen, I. Reid, and M. Milford, “Deep learning features at scale for visual place recognition,” in *Proceedings - IEEE International Conference on Robotics and Automation*, 2017.

[7] R. Gomez-Ojeda, M. Lopez-Antequera, N. Petkov, and J. Gonzalez-Jimenez, “Training a Convolutional Neural Network for Appearance-Invariant Place Recognition,” in *ArXiv pre-print*, 2015.

[8] D. K. Schroff, Florian and J. Philbin, “”Facenet: A unified embedding for face recognition and clustering.” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015.