Learning Multi-scene Absolute Pose Regression with Transformers Supplementary Materials

Yoli Shavit Ron Ferens Yosi Keller
Faculty of Engineering, Bar Ilan University, Ramat-Gan, Israel
{yolisha, ronferens, yosi.keller}@gmail.com

Contents

1. Appendix

1.1. Attention and Transformers

Attention mechanisms [1] are neural network layers that aggregate information from the entire input sequence. The aggregation is often computed by a sequence-to-sequence architecture, where the inner-products (interactions) between the two sequences are used to compute the aggregation weights. Attention models consist of an Encoder and Decoder. The Encoder implements self-attention that maps the input sequence into a higher dimensional space, that is fed into the Decoder alongside a query sequence, outputting the result sequence. Attention allows to numerically emphasize the contribution of the task-informative image locations, in contrast to the visual clutter. Transformers were introduced by Vaswani et al. [23] as a novel formulation of attention-based Encoders and Decoders for sequence encoding that does not use RNN layers such as LSTM and GRU. Transformers consist of multiple stacked Multi-Head Attention and Feed Forward layers. As no recurrent layers are used, the relative position and sequential order of the sequence elements are induced by adding positional encodings to the embedded representation. Transformers were shown to outperform RNNs in encoding long sequences, and were applied in multiple recent works in natural language processing (NLP) [9, 18] and computer vision [7, 10]. In this work, we propose a hybrid CNN-Transformer architecture, inspired by recent advancements in visual transformers for multi-object detection [7]. We employ encoders to adaptively aggregate activation maps for position and orientation regression and use decoders to decode aggregated representations with respect to query scenes encoding.

1.2. Data Augmentation and Training

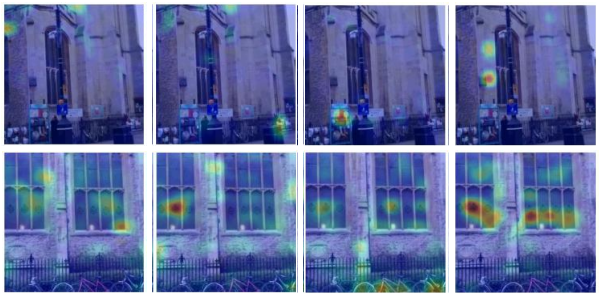
We follow the augmentation procedure described in [13]. During training, images are first rescaled so that the smaller edge is resized to 256 pixels and then randomly cropped to a 224×224 sized image. In addition, the brightness, contrast and saturation are randomly jittered. At test time, the center crop is taken after rescaling without any further augmentations. For the 7Scenes dataset we train with the aforementioned augmentation scheme for 30 epochs and reduce the learning rate by half every 10 epochs. For the Cambridge-Landmarks dataset, we first train without augmentations for 550 epochs, reducing the learning rate by half every 200 epochs. We then train for another 40 epochs (following the same learning rate decay regime) with augmentations, optimizing only the position branch (freezing all other weights) with the position loss described in Eq. 6 in the main text. Since this dataset also presents large variations in scene size we sample images from smaller scenes more frequently to achieve a data equalization effect.

1.3. Visualization of Decoder Attention

In order to gain additional insights into the scene-specific features learned by our model, we further visualize the attentions $\{\mathbf{X}_i\}_1^N$ at the outputs of the positional decoder two images from the St. Mary scene. In addition, we measure and rank the decoder outputs by summing over the corresponding attention map. Indeed, the strongest response is obtained with the output corresponding to this scene (Fig. 1d). Interestingly, the activations related to the ShopFacade (Fig. 1c) scene attend to the lower part of the input images, which is typically includes the key features in images from this scene.

1.4. Model Scalability

Two of the main benefits of single-scene APRs, compared to other classes of localization methods, are their runtime (10ms [3]) and memory footprint. However, in order to cover a site with N scenes, N models need to be stored and



(a) Kings C. (b) Old Hospital (c) Shop Facade (d) St. Mary

Figure 1: Translational Decoder attention visualization $\{\mathbf{X}_i\}_1^N$. Each activation relates to a different scene. The activations are due to two input images from the St Marys scene. The activations of the corresponding scene are notably stronger.

selected from during inference time. For example, serving a site of 1000 scenes with the PoseNet model [13] requires 1000×50Mb = 5Gb. In order to evaluate the scalability of our model, we measure its runtime and memory footprint with an increasing number of scenes. For this purpose, we instantiated models for an increasing number of scenes, and measured their memory signature and the runtime of their forward pass, allowing us to evaluate the scalability of our approach in the absence of real data. The results of this experiment are shown in Table 1. We compare two variants of our model: the architecture used for comparative analysis (Section 4.2 in the main text), with six layers for each encoder and decoder, and a shallower model with two layers per encoder/decoder, for which we report similar performance as part of our ablation study (Section 4.4 in the main text). The memory footprint of our model remains relatively constant, where increasing from 4 to 1000 scenes adds only 2Mb. In addition, both variants require under 80Mb, before any optimization. The runtime of the model remains constant in the range of 4-100 scenes and increases by $\sim 1.5x$ for 1000 scenes. Assuming a constant selection time, our model is $2 - 5x$ slower compared to a single scene APR. This can be expected due to the runtime complexity of the MHA operation, which is quadratic with respect to the sequence length (number of scenes). However, a significant acceleration can be obtained with recent linear-time MHA formulations [8] and other general optimization methods [22] in order to enable competitive runtime.

1.5. Related Work

Although our works focuses standalone light-weight end-to-end camera pose regressors, we include cross-comparison of representative methods from different localization approaches on the Cambridge Landmarks and the 7Scenes datasets (Tables 2 and 3 respectively).

Table 1: Runtime (in ms) and memory footprint (in Mb) as a function of the number of learned scenes. We show the results for two instantiations of our model, using two or six layers for all encoders and decoders. highlighted in bold.

Num. Scenes	**Runtime [ms]**		**Memory [Mb]**	
Num. Layers	2	6	2	6
1	18.8	34.6	40.8	74.6
4	18.8	35	40.8	74.6
7	19.2	35.2	40.8	74.6
10	19.2	35.2	40.8	74.6
100	19.6	35.4	41.0	74.8
500	21.0	41.0	41.8	75.6
1000	27.0	58.6	42.8	76.7

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[2] Vassileios Balntas, Shuda Li, and Victor Prisacariu. Relocnet: Continuous metric learning relocalisation using neural nets. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[3] Hunter Blanton, Connor Greenwell, Scott Workman, and Nathan Jacobs. Extending absolute pose regression to multiple scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 38–39, 2020.

[4] E. Brachmann, A. Krull, S. Nowozin, J. Shotton, F. Michel, S. Gumhold, and C. Rother. Dsac â€" differentiable ransac for camera localization. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2492–2500, Los Alamitos, CA, USA, jul 2017. IEEE Computer Society.

[5] E. Brachmann and C. Rother. Learning less is more - 6d camera localization via 3d surface regression. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4654–4662, 2018.

[6] Samarth Brahmbhatt, Jinwei Gu, Kihwan Kim, James Hays, and Jan Kautz. Geometry-aware learning of maps for camera localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 213–229, Cham, 2020. Springer International Publishing.

Table 2: Comparative analysis of pose regressors on the Cambridge Landmarks dataset (outdoor localization). We report the median position/orientation error in meters/degrees for each method. Best results are highlighted in bold.

Method	K. College	Old Hospital	Shop Facade	St. Mary
PoseNet [13]	1.92/5.40	2.31/5.38	1.46/8.08	2.65/8.48
Dense PoseNet [13]	1.66/4.86	2.57/5.14	1.41/7.18	2.45/7.96
Bayesian [11]	1.74/4.06	2.57/5.14	1.25/7.54	2.11/8.38
LSTM-Pose [24]	0.99/3.65	1.51/4.29	1.18/7.44	1.52/6.68
SVS-Pose [17]	1.06/2.81	1.50/4.03	0.63/5.73	2.11/8.11
PoseNet (Reproj) [12]	0.99/1.10	2.17/2.9	1.05/4.0	1.49/3.40
VLocNet [21]	0.836/1.42	1.07/2.411	0.593/3.53	0.631/3.91
DSAC [4]	0.30/0.5	0.33/0.6	0.09/0.40	0.55/1.6
DSAC++ [5]	0.18/0.3	0.20/0.3	0.06/0.30	0.13/0.4
Active Search [20]	0.42/0.6	0.44/1.0	0.12/0.40	0.19/0.5

Table 3: Comparative analysis of pose regressors on the 7Scenes dataset (indoor localization). We report the median position/orientation error in meters/degrees for each method. Best results are highlighted in bold.

Method	Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs
PoseNet [13]	0.32/8.12	0.47/14.4	0.29/12.0	0.48/7.68	0.47/8.42	0.59/8.64	0.47/13.8
Dense PoseNet [13]	0.32/6.60	0.47/14.0	0.30/12.1	0.48/7.24	0.49/8.12	0.58/8.34	0.48/13.8
Bayesian [11]	0.37/7.24	0.43/13.7	0.31/12.0	0.48/8.04	0.61/7.08	0.58/7.54	0.48/13.1
LSTM-Pose [24]	0.24/5.77	0.34/11.9	0.21/13.7	0.30/8.08	0.33/7.00	0.37/8.83	0.40/13.7
Hourglass-Pose [16]	0.15/6.53	0.27/10.84	0.19/11.63	0.21/8.48	0.25/7.01	0.27/10.84	0.29/12.46
BranchNet [25]	0.18/5.17	0.34/8.99	0.20/14.15	0.30/7.05	0.27/5.10	0.33/7.40	0.38/10.26
PoseNet (Reproj) [12]	0.13/4.48	0.27/11.3	0.17/13.0	0.19/5.55	0.26/4.75	0.23/5.35	0.35/12.4
MapNet [6]	0.09/3.24	0.20/9.29	0.12/8.45	0.19/5.45	0.19/3.96	0.20/4.94	0.27/10.57
VLocNet [21]	0.036/1.71	0.039/5.34	0.046/6.64	0.039/1.95	0.037/2.28	0.039/2.20	0.097/6.48
VLocNet++ [19]	0.023/1.44	0.018/1.39	0.016/0.99	0.024/1.14	0.024/1.45	0.025/2.27	0.021/1.08
DGRNets [15]	0.016/1.72	0.011/2.19	0.017/3.56	0.024/1.95	0.022/2.27	0.018/1.86	0.017/4.79
NNnet [14]	0.13/6.46	0.26/12.72	0.14/12.34	0.21/7.35	0.24/6.35	0.24/8.03	0.27/11.80
RelocNet [2]	0.12/4.14	0.26/10.4	0.14/10.5	0.18/5.32	0.26/4.17	0.23/5.0	0.28/7.53
DSAC [4]	0.02/1.2	0.04/1.5	0.03/2.7	0.04/1.6	0.05/2.0	0.05/2.0	1.17/33.1
DSAC++ [5]	0.02/0.5	0.02/0.9	0.01/0.8	0.03/0.7	0.04/1.1	0.04/1.1	0.09/2.6
Active Search [20]	0.04/2.0	0.03/1.5	0.02/1.5	0.09/3.6	0.08/3.1	0.07/3.4	0.03/2.2

[8] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020.

[11] Alex Kendall and Roberto Cipolla. Modelling uncertainty in deep learning for camera relocalization. In Danica Kragic, Antonio Bicchi, and Alessandro De Luca, editors, *2016 IEEE International Conference on Robotics and Automation, ICRA 2016, Stockholm, Sweden, May 16-21, 2016*, pages 4762–4769. IEEE, 2016.

[12] A. Kendall and R. Cipolla. Geometric loss functions for camera pose regression with deep learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6555–6564, 2017.

[13] A. Kendall, M. Grimes, and R. Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2938–2946, 2015.

[14] Z. Laskar, I. Melekhov, S. Kalia, and J. Kannala. Camera relocalization by computing pairwise relative poses using convolutional neural network. In *2017 IEEE International*

Conference on Computer Vision Workshops (ICCVW), pages 920–929, 2017.

[15] Yimin Lin, Zhaoxiang Liu, Jianfeng Huang, Chaopeng Wang, Guoguang Du, Jinqiang Bai, and Shiguo Lian. Deep global-relative networks for end-to-end 6-dof visual localization and odometry. In *Pacific Rim International Conference on Artificial Intelligence*, pages 454–467. Springer, 2019.

[16] Iaroslav Melekhov, Juha Ylioinas, Juho Kannala, and Esa Rahtu. Image-based localization using hourglass networks. In *2017 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2017, Venice, Italy, October 22-29, 2017*, pages 870–877. IEEE Computer Society, 2017.

[17] Tayyab Naseer and W. Burgard. Deep regression for monocular camera-based 6-dof global localization in outdoor environments. *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1525–1530, 2017.

[18] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[19] N. Radwan, A. Valada, and W. Burgard. Vlocnet++: Deep multitask learning for semantic visual localization and odometry. *IEEE Robotics and Automation Letters*, 3(4):4407–4414, 2018.

[20] T. Sattler, B. Leibe, and L. Kobbelt. Efficient effective prioritized matching for large-scale image-based localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1744–1756, 2017.

[21] Abhinav Valada, Noha Radwan, and Wolfram Burgard. Deep auxiliary learning for visual localization and odometry. *ICRA*, pages 6939–6946, 2018.

[22] Han Vanholder. Efficient inference with tensorrt, 2016.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc., 2017.

[24] F. Walch, C. Hazirbas, L. Leal-Taixé, T. Sattler, S. Hilsenbeck, and D. Cremers. Image-based localization using lstms for structured feature correlation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 627–637, 2017.

[25] J. Wu, L. Ma, and X. Hu. Delving deeper into convolutional neural networks for camera relocalization. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5644–5651, 2017.