Supplementary Materials– Joint Spatial and Layer Attention for Convolutional Networks

Tony Joseph[1]
tony.joseph@uoit.net
Konstantinos G. Derpanis[23]
www.scs.ryerson.ca/kosta
Faisal Z. Qureshi[1]
faculty.uoit.ca/qureshi

[1] Faculty of Science
Ontario Tech University
Oshawa, Canada

[2] Department of Computer Science
Ryerson University
Toronto, Canada

[3] Samsung AI Centre
Toronto, Canada

1 Detailed Attention Architecture

Figure 1, illustrates the layer selection mechanism. The mechanism receives input h_t from ConvLSTM. It then performs an average pool and an intermediate gate embedding before prediction. We add the Gumbel samples to the predicted logits and perform an *argmax* to select the optimal layer. The gate embedding layer dimension E is much smaller than C. This gate embedding layer helps build a possible representation of incoming features at every LSTM steps, without significantly increasing the network parameters.

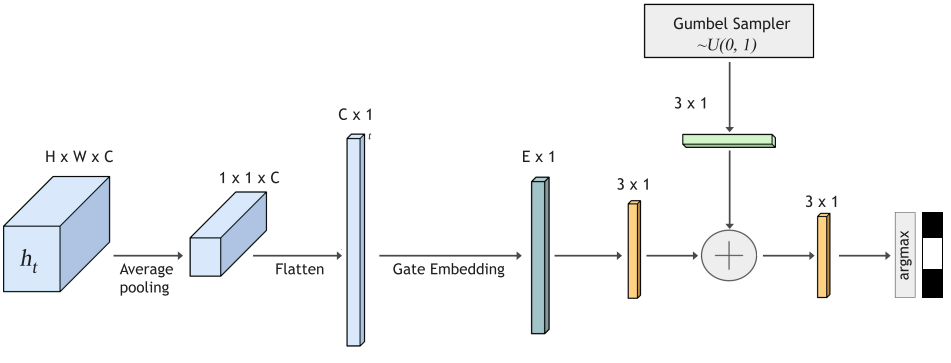

Figure 1: Layer Selection Mechanism.

Figure 2, illustrates the soft attention mechanism. Unlike the soft attention mechanism proposed in Xu *et al*. [11] our's replace fully-connected layers with convolutional layers. Specifically, we used multi-convolutional layers that uses different kernel sizes similar to an inception module. At each time step t, the module receives h_t from ConvLSTM and the selected feature layer F_t. The ConvLSTMs hidden state h_t is first converted to the appropriate

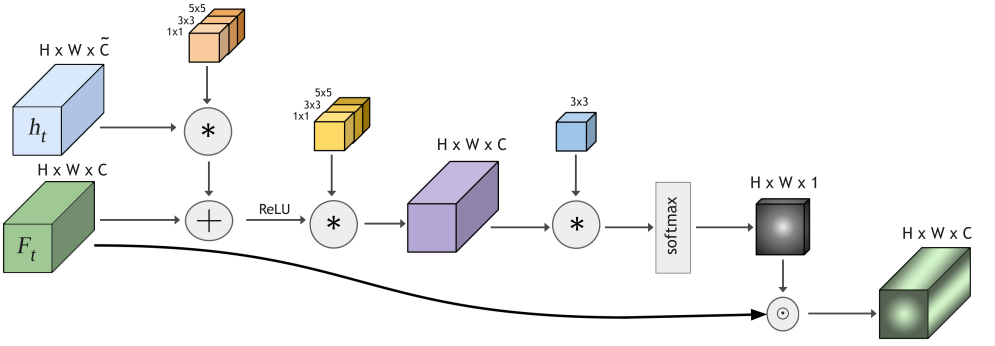

Figure 2: Soft Attention Mechanism.

channel size of the feature map. We add the embedding h_t and feature layer F_t. Then we apply a non-linearity (Leaky ReLU). After which we compute the attention weights and apply softmax to get the attention map. Then an element-wise multiplication is performed between features and attention map to get the final output of the soft attention module. The Multi-ConvLSTM is applied to attention output. At each time step the LSTM output is used for prediction. In Section suggest convolutional attention and LSTMs yield better results. We did try using fully-connected LSTMs; however, the system consistently failed to pick different locations in the image during successive LSTM steps.

2 Datasets

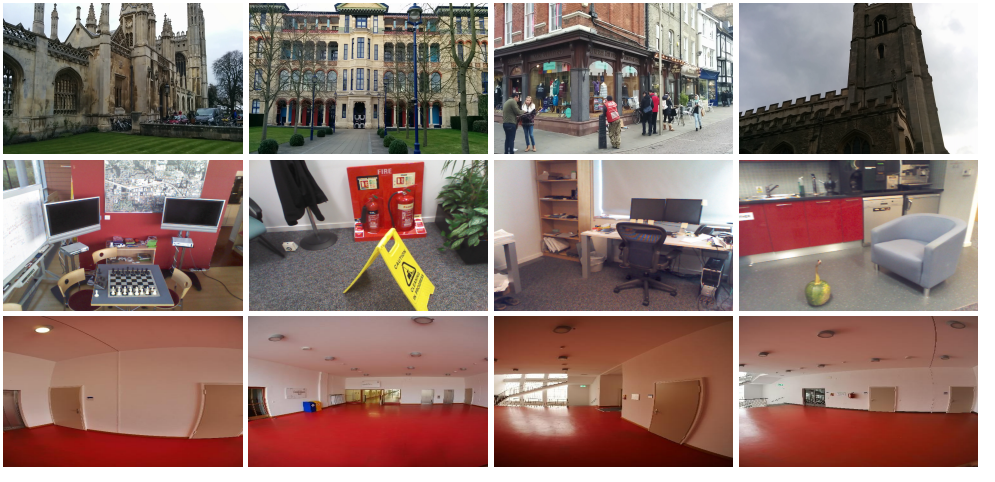

Figure 3: (a) Top row: Cambridge Landmarks Dataset. King's College, Old Hospital, Shop Facade and St. Mary's Church. (b) Middle row: 7-Scenes (subset). Chess, Fire, Office and Pumpkin. (c) Bottom row: TUM-LSI.

Cambridge Landmarks [9] A large scale outdoor dataset, containing five outdoor datasets. For our experiments, we only use the four datasets that were used by [4] and [9]. The dataset consists of RGB images. Six degrees-of-freedom camera poses are provided for each image.

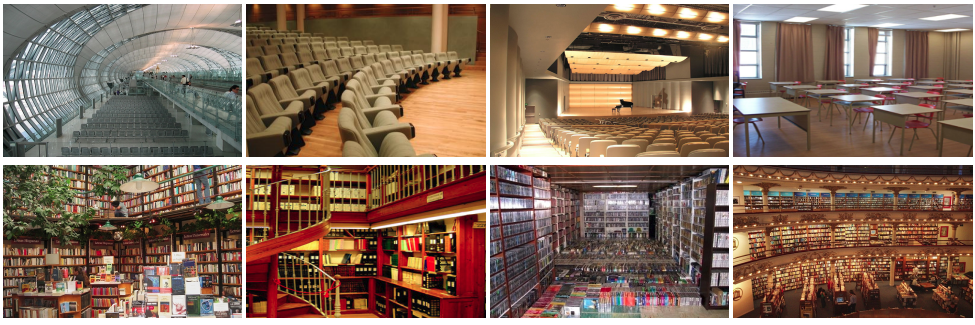

Figure 4: MIT-67 Indoor Scene Dataset. (a) Top row: Airport, Auditorium, Concert Hall and Classroom. A network can have a hard time classifying them by just focusing on specific properties, since all of them contain large hallways with chairs. (b) Bottom row: Bookstore, Library, Video Store and Library. This set of images have almost the same structure and objects which makes these scenes very ambiguous.

The dataset was collected using a smart phone, and structure from motion was employed to label each image with its corresponding camera pose.

7-Scenes [7] A small scale indoor dataset, which consists of seven different scenes. These scenes were obtained using Kinect RGB-D camera, and KinectFusion[2] was used to obtain the ground truth. We use the train/test split used by [4] and [9]. Scene contain ambiguous regions, which makes camera localization difficult.

TU Munich Large-Scale Indoor (TUM-LSI) [9] An indoor dataset, which covers an area of two orders of magnitude larger than that covered by the 7Scenes dataset. It consists of 875 training images and 220 testing images. We use the train/test split used by [9]. This is a challenging dataset to localize due to repeated structural elements with nearly identical appearance.

MIT-67 indoor scenes [8] Images taken primarily in four different indoor environments—store, home, public spaces, leisure and working places. The dataset contains 67 categories in total. We used the official train/test split provided by [8]. Each category has 80 training images and 20 testing images.

3 Extended Implementation details

For both camera pose estimation and indoor scene classification, we used the same pre-trained CNN layers as used by previous methods. We used the original GoogLeNet weights trained on Places[1] [11]. By necessity, we converted these provided trained network weights to be able to use these in *TensorFlow*. The batch size during training was set to 40. The intial memory states of the LSTM (Memory state c_0 and Hidden state h_0) is typically set to zero. Similar to [10], we learn the the initial states.The ConvLSTM hidden size is set to 96.

3.1 Multi-Convolutional Approach

In this section, we describe our motivation for using the multi-convolutional approach. To showcase how we arrived at the proposed approach, we provide evaluation on all three

[1] http://places.csail.mit.edu/downloadCNN.html

Dataset	PoseNet [9]	LSTM-PoseNet [9]	Ours	
			Convolutional Spatial Attention	Improvement (meter, degree) %
King's College	1.66 m, 4.86°	**0.99 m**, 3.65°	1.39 m, **2.63**°	-27.2, +27.6
Old Hospital	2.62 m, 4.90°	**1.51 m**, 4.29°	3.72 m, **4.24**°	-120.5, +6.9
Office	0.48 m, 7.24°	**0.30 m**, 8.08°	0.64 m, **7.89**°	-103.3,+3.2
Stairs	0.48 m, 13.1°	**0.40 m**, 13.7°	0.48 m, **12.8**°	-15.0, +6.5
TUM-LSI	1.87 m, 6.14°	**1.31 m**, 2.79°	3.93 m, **2.15**°	+16, +22.9

Table 1: Median localization error achieved by the convolutional attention model on a subset of camera pose estimation datasets: Cambridge Landmarks, 7-Scenes, and TUM-LSI dataset. Bold values indicate the lowest error achieved for each row.

Dataset	PoseNet [9]	LSTM-PoseNet [9]	Ours	
			Multi-Conv. Spatial Attention	Improvement (meter, degree) %
King's College	1.66 m, 4.86°	0.99 m, **3.65**°	**0.95** m, 4.11°	+4.04, -12.6
Old Hospital	2.31 m, 5.38°	**1.51** m, **4.29**°	1.76 m, 4.44°	-16.5, -3.49
Office	0.48 m, 7.24°	0.30 m, 8.08°	**0.28** m, **7.52**°	+6.67, +6.93
Stairs	0.48 m, 13.1°	0.40 m, 13.7°	**0.32** m, **12.7**°	+20.0, +9.40
TUM-LSI	1.87 m, 6.14°	1.31 m, **2.79**°	**1.12** m, 3.66°	+14.5, -2.88

Table 2: Median localization error achieved by the multi-convolutional attention model on a subset of camera pose estimation datasets: Cambridge Landmarks, 7-Scenes, and TUM-LSI dataset. Bold values indicate the lowest error achieved for each row.

datasets for the pose estimation. We initially started with the same implementation as Xu *et al*. [10] for soft attention, by using fully connected layers. The model ended up overfitting the data and showed poor performance on the test set. Also, the network converged to select only a single spatial feature instead of probing through the other spatial features at different LSTM time-steps. Our first solution was converting fully connected layers into fully convolutional layers. The results for this approach on pose estimation is shown in Table 1. The results shown is quite far from [9] especially on the position, but interestingly error was close to [4].

We found that our model was underfitting the training data. Naively increasing the depth size or kernel size was not showing any significant improvements. Therefore by taking inspiration from the inception module proposed in GoogLeNet [8], we converted each convolutional layer into multi-convolutional layers. We used three convolutional kernels with kernel sizes of 1x1, 3x3 & 5x5 and stacked their final output together. Similarly, in the case of ConvLSTM, we used four convolutional kernels with kernel sizes of 1x1, 3x3, 5x5 & 7x7. Then stacked their final output together for prediction. This approach helped improve results significantly as shown in Table 2. After which we applied our contribution of layer selection

mechanism to form layer-spatial attention. The final results for pose estimation is shown in Table 1 in the main paper.

4 Extended results

4.1 Results for Manual Layer Search

In this section, we show an extensive list of classes in MIT-67 indoor scene classification dataset. This table is an extension to the Table 3 from the main main paper. This is provided to showcase how different layers of CNN capture distinctive information that can help further improve the result.

Scene	Layer 3B	Layer 4E	Layer 5B
Office	33.3	**52.3**	42.8
Library	**65.0**	45.0	60.0
Wine Cellar	71.4	**76.1**	61.9
Fastfood Restaurant	58.8	**88.2**	70.5
Operating Room	47.3	**52.6**	36.8
Train Station	**85.0**	65.0	60.0
Airport-inside	40.0	60.0	**75.0**
Closet	77.7	88.8	**94.4**
Game Room	45.0	75.0	**80.0**
Garage	72.2	77.7	**94.4**
Dining room	38.8	66.6	**77.7**
Locker room	66.6	85.7	**100.0**

Table 3: Indoor scene classification. Mean Accuracy results (%) after applying spatial soft attention to feature maps from different GoogLeNet layers. Top rows show the classes that improve as we look at different layers. Bottom rows show the classes that decrease performance when looking at other layers. Bold values indicate the highest accuracy achieved for each row.

4.2 Results for five Conv-LSTM steps

Dataset	Area or Volume	PoseNet [■]	Bayesian PoseNet [■]	LSTM PoseNet [■]	Ours					
					Conv-LSTM Step-1	Conv-LSTM Step-2	Conv-LSTM Step-3	Conv-LSTM Step-4	Conv-LSTM Step-5	Improvement (meter, degree)
Old Hospital	2000 m^2	2.62 m, 4.90°	2.57 m, 5.14°	1.51 m, 4.29°	1.62 m, 4.11°	1.51 m, 4.02°	**1.36** m, **3.95**°	1.55 m, 4.46°	1.64 m, 4.20°	+9.93, +7.92
St. Marys Church	4800 m^2	2.45 m, 7.96°	2.11 m, 8.38°	1.52 m, 6.68°	1.62 m, 7.22°	1.59 m, 5.94°	**1.42** m, **6.07**°	1.49 m, 5.87°	1.58 m, 6.51 °	+6.57, +1.64
Office	7.5 m^3	0.48 m, 7.24°	0.48 m, 8.04°	0.30 m, 8.08°	0.29 m, 7.63°	0.29 m, 7.23°	**0.29** m, **8.02**°	0.29 m, 8.07°	0.30 m, 8.12 °	+3.33, +0.74
Stairs	7.5 m^3	0.48 m, 13.1°	0.48 m, 13.1°	0.40 m, 13.7°	0.32 m, 9.98°	0.31 m, 10.5°	**0.29** m, **12.0**°	0.31 m, 12.0°	0.33 m, 10.9 °	+27.5, +12.4
TUM-LSI	5575 m^2	1.87 m, 6.14°	-	1.31 m, 2.79°	1.32 m, 3.82°	1.26 m, 3.69°	**0.98** m, **2.74**°	1.14 m, 3.33°	1.18 m, 3.68 °	+25.1, +1.79

Table 4: Median localization error achieved by our proposed attention model over five-time steps on subset of Cambridge Landmarks, subset of 7-Scenes, and TUM-LSI. Bold values indicate the lowest error achieved for each row. Improvement is reported with respect to LSTM-PoseNet [9].

CNNaug-SVM [■]	S^2ICA [■]	GoogLeNet [■]	**Ours**					
			Conv-LSTM Step-1	Conv-LSTM Step-2	Conv-LSTM Step-3	Conv-LSTM Step-4	Conv-LSTM Step-5	Improvement (%)
69.0 %	71.2 %	73.7 %	74.5 %	**77.1** %	76.0 %	75.4	74.8	+3.4

Table 5: Mean accuracy results for indoor scene classification on MIT-67. The proposed method achieves the highest accuracy (shown in boldface). Improvement is reported with respect to the GoogLeNet [8] baseline.

Camera localization. We did an experimental study for a subset of scenes from camera localization dataset shown in Table 4. We concluded that for the camera position estimation Conv-LSTM step three on average provides the best result.

Indoor Scene Classification. We did an experimental study on MIT-67 indoor scene, shown in Table 5. We concluded that for the Indoor Scene Conv-LSTM step two on average provides the best result.

References

[1] Munawar Hayat, Salman H Khan, Mohammed Bennamoun, and Senjian An. A spatial layout and scale invariant feature representation for indoor scene classification. *Proc. of the IEEE Transactions on Image Processing*, pages 4829–4841, August 2016.

[2] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proc. of the ACM symposium on User interface software and technology*, pages 559–568, 2011.

[3] Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, pages 5574–5584, December 2017.

[4] Alex Kendall, Matthew Grimes, and Roberto Cipolla. PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2938–2946, June 2015.

[5] Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *Proc. of the IEEE Conference on Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 413–420, June 2009.

[6] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. CNN features off-the-shelf: An astounding baseline for recognition. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition workshops (CVPRW)*, pages 806–813, June 2014.

[7] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in RGB-D images. In *Proc. of the IEEE Conference on Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2930–2937, June 2013.

[8] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proc. of the IEEE Conference on Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, June 2015.

[9] Florian Walch, Caner Hazirbas, Laura Leal-Taixé, Torsten Sattler, Sebastian Hilsenbeck, and Daniel Cremers. Image-based localization using LSTMs for structured feature correlation. In *Proc. of the IEEE Conference on International Conference on Computer Vision (ICCV)*, pages 627–637, October 2017.

[10] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proc. of the International Conference on Machine Learning, (ICML)*, pages 2048–2057, July 2015.

[11] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, pages 487–495, December 2014.