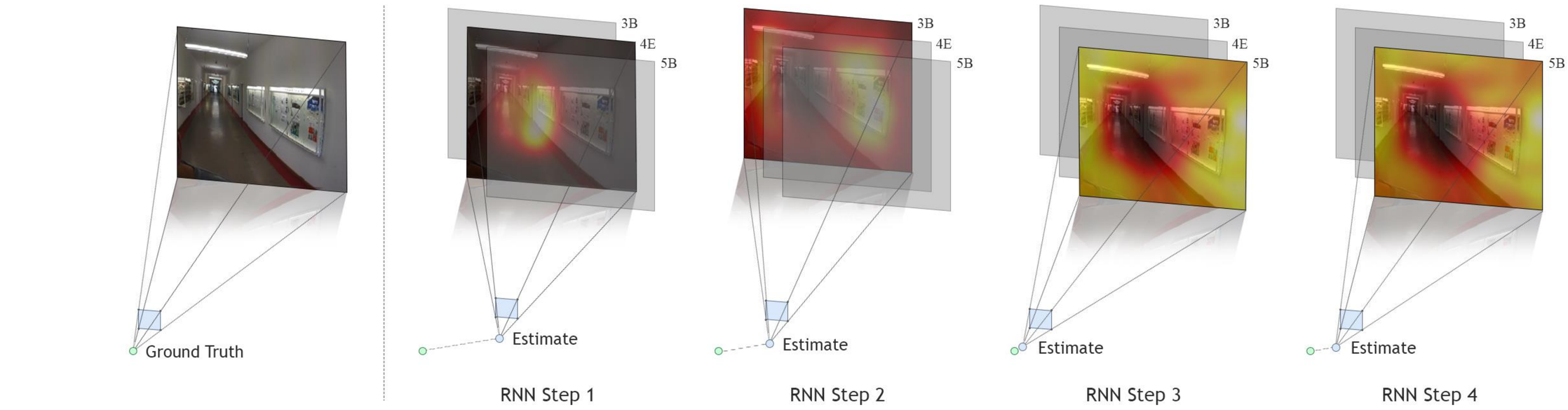
Joint Spatial and Layer Attention for Convolutional Networks

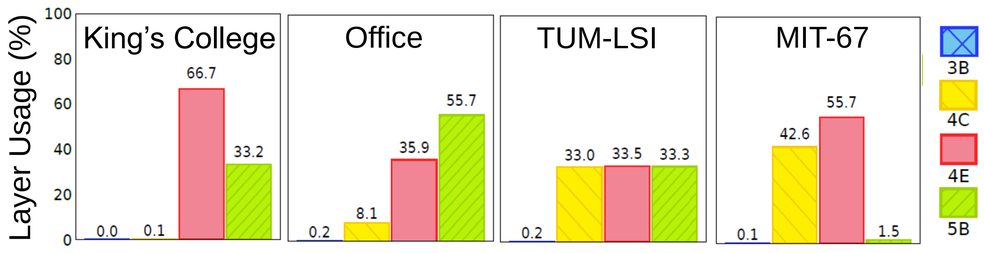
Tony Joseph, Konstantinos G. Derpanis, Faisal Z. Qureshi

Overview

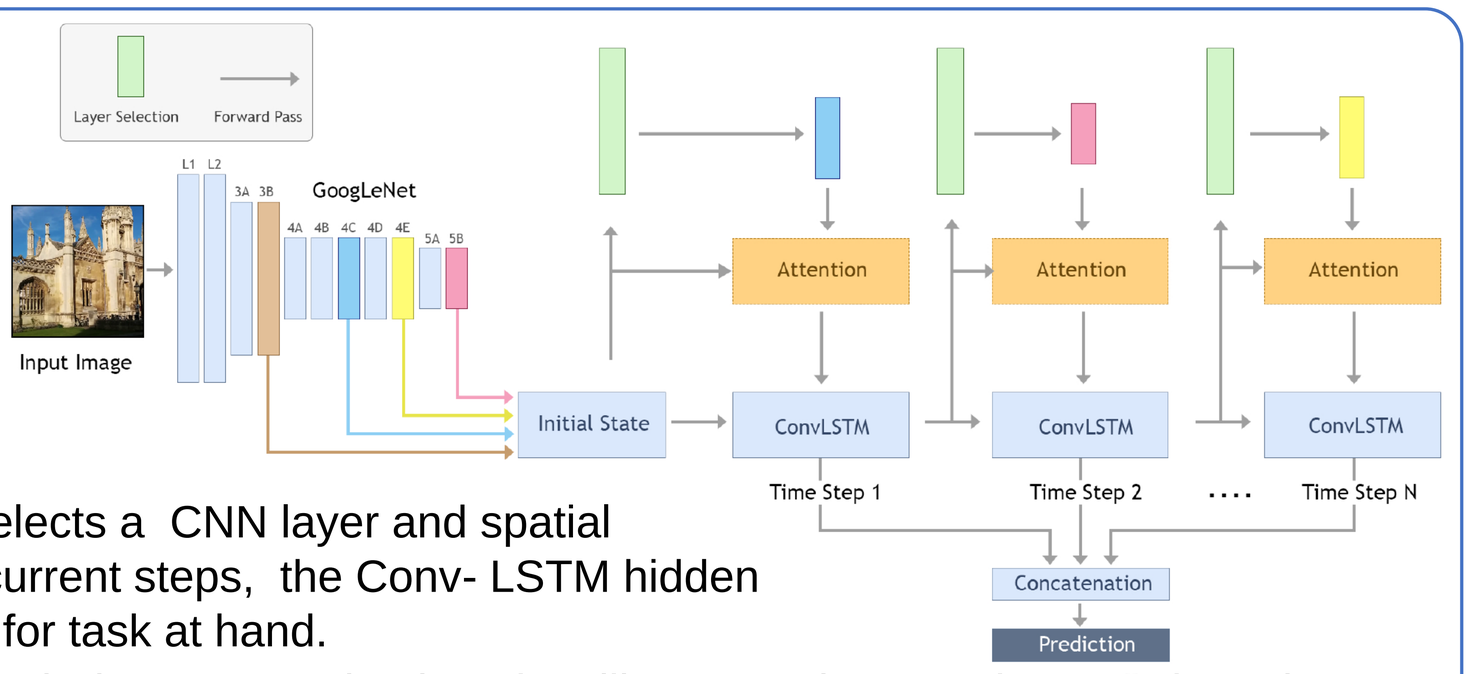


Overview of our approach to 6-DoF camera localization. Given a set of CNN feature layers (GoogleNet Conv-[3B, 4E, 5B] layers shown) our approach to attention uses a RNN to sequentially select a set of feature layers (highlighted by the non-grey images) and corresponding locations in the layers (highlighted by the heat maps, where darker regions indicate higher spatial weight). Finally, the processed attended features are used for regressing the camera position and orientation.

Approach



Layer Selection Frequencies on the subset of datasets on the test set. The bins refer to the GoogLeNet Conv-{3B, 4C, 4E, 5B} layers. The vertical axis represents layer usage percentages.



- At each Conv-LSTM step, the layer attention selects a CNN layer and spatial attention localizes a region within it. After N recurrent steps, the Conv- LSTM hidden states for all steps are concatenated and used for task at hand.
- The Layer Selection Frequencies shows that each datasets predominately utilize more than one layer. Selected layers differs widely amongst the datasets.

Quantitative Results

We evaluated our approach on:

1. Camera Localization (regression task)
2. Indoor Scene classification (classification task)

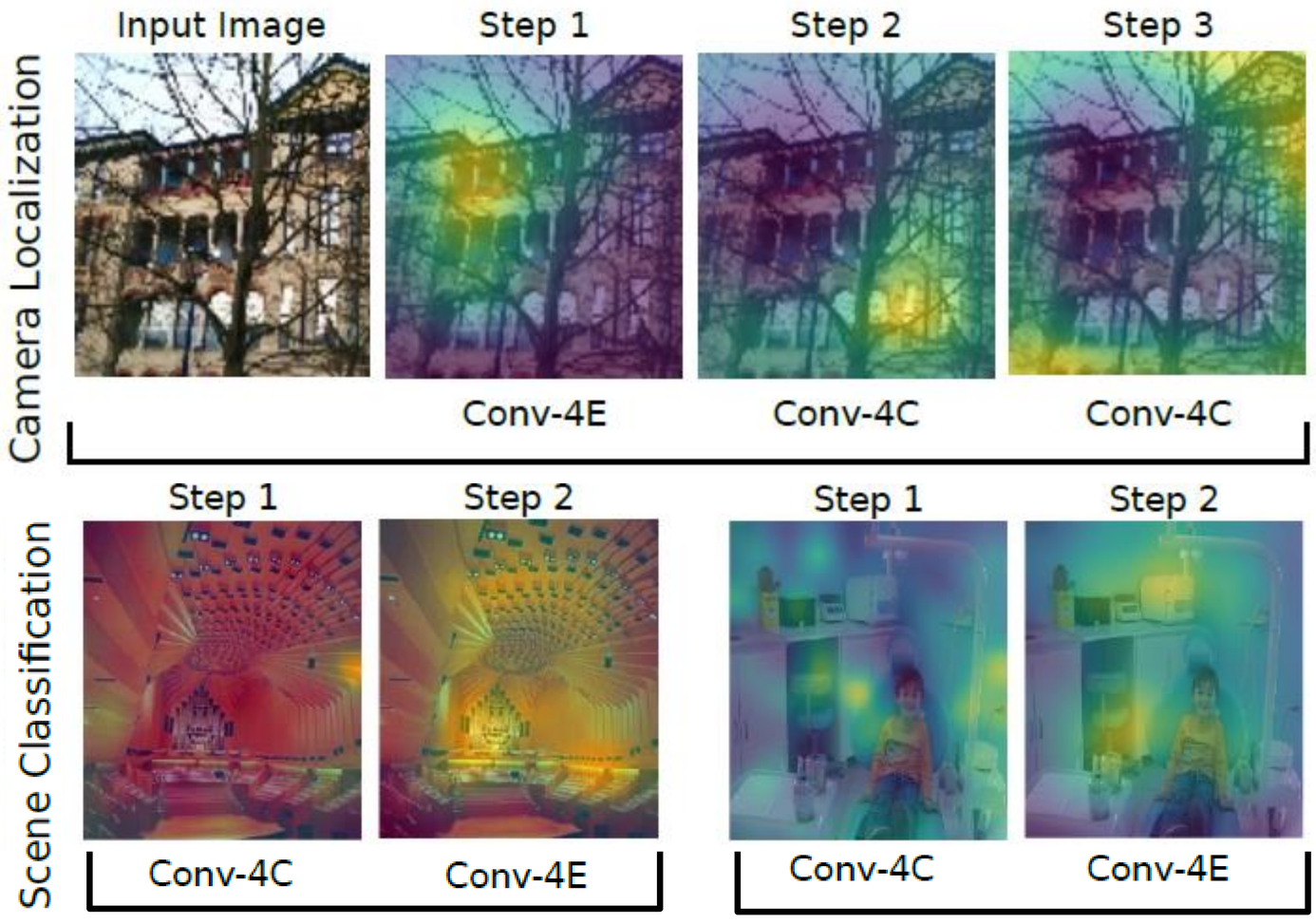
Camera localization
(Median Error results reported with respect to [2])

Dataset	Position (Improvement %)	Orientation (Improvement %)
Cambridge Landmarks	12.3	13.9
7-Scenes	19.3	8.83
TUM-LSI	25.1	1.79

Indoor Scene Classification
(Mean Accuracy Improvement (%) with respect baseline [1])

MIT-67	3.4

Qualitative Results



Acknowledgements

1. Konstantinos Derpanis and Faisal Qureshi are both supported by Canadian NSERC discovery grant.
2. NVIDIA corporation for supporting this research with the donation of a Titan-Xp GPU.

References

1. Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir, Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions, CVPR-2015.
2. Florian Walch, Caner Hazirbas, Laura Leal-Taixé, Torsten Sattler, Sebastian Hilsenbeck, and Daniel Cremers. Image-based localization using LSTMs for structured feature correlation, ICCV-2017.
3. 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, ICML-2015.