CS-VQA: Visual Question Answering with Compressively Sensed Images

Li-Chi Huang†, Kuldeep Kulkarni, Anik Jha†, Suhas Lohit†, Suren Jayasuriya†#, Pavan Turaga†#*

† School of Electrical, Computer and Energy Engineering, ASU # School of Arts, Media and Engineering, ASU

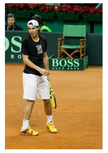
** College of Electrical and Computer Engineering, CMU*

(This work has been accepted in International Conference on Image Processing (ICIP) 2018)

Geometric Media Lab

Introduction

- The Visual Question Answering (VQA) problem deals with the task of answering an open ended question posed with respect to a given image
- We explore whether the underlying representation of visual data in 2D images is even critical for VQA performance
- In particular, ability to use sub-Nyquist rate sensed measurements of natural images in a VQA architecture can help adapting VQA techniques to resource-constrained platforms like HoloLens



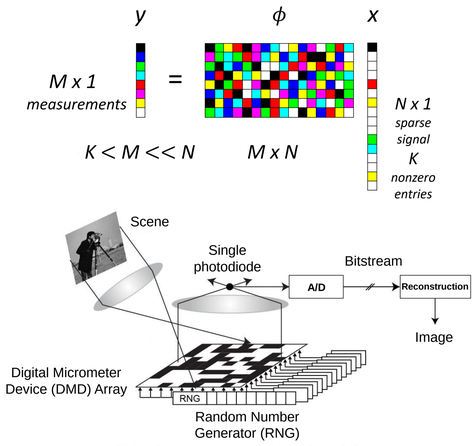
Question: Are these shoes appropriate for playing tennis?

Answer: Yes

Question: Are these zebras on a road?

Answer: Yes

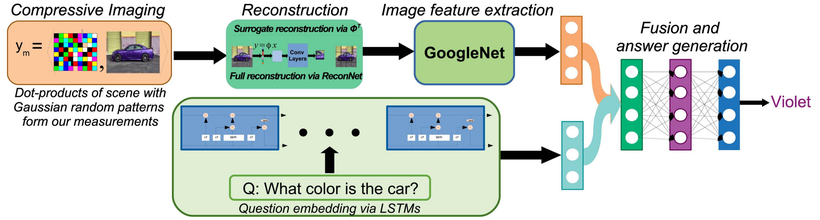
Compressive Sensing (CS)



This picture has been taken from [1]

- In Compressive Imaging (CI), random projections of the signal are directly acquired without first collecting the pixels/voxels
- CS imaging decreases energy consumption and bandwidth utilization
- The Single-Pixel Camera (SPC)[1] is a popular example of a compressive imager
- Much of CS theory has attempted to reconstruct the signals/images with different assumptions on image statistics [1,2]

CS-VQA Architecture



Experimental Results

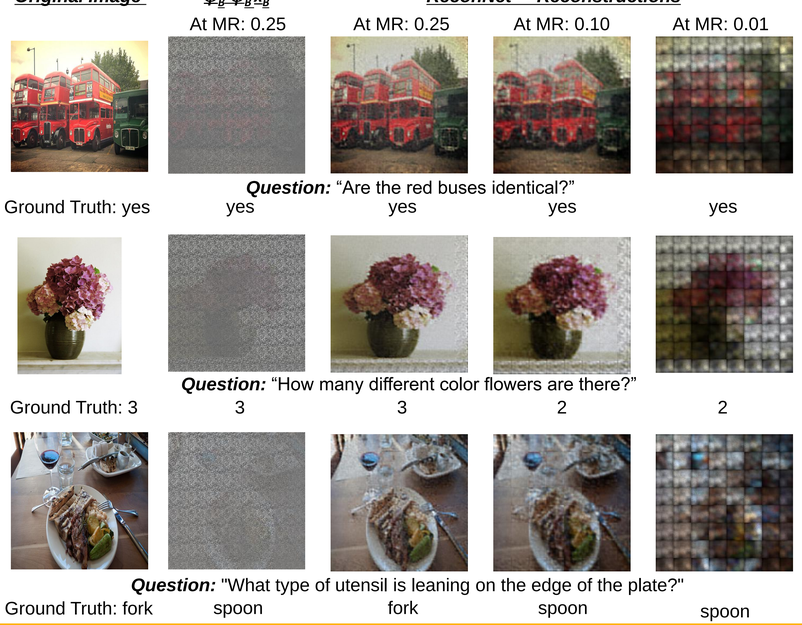
Open-ended VQA v1.0[4] results with various CS reconstructions, and their corresponding accuracy(%)

CS Reconstruction	All	Yes/No	Number	Other
$\phi_B^T \phi_B x_B$ ***(MR = 0.25)***	52.98	79.50	33.03	38.15
ReconNet (MR = 0.25)	54.22	79.85	33.28	40.21
ReconNet (MR = 0.10)	51.40	79.13	33.20	35.21
ReconNet (MR = 0.01)	51.05	78.77	32.92	34.87
Oracle VQA v1.0				
LSTM + VGG	57.75	80.50	36.77	43.08
Question Only	50.39	78.41	34.68	30.03

Open-ended VQA v2.0 [5] results with various CS reconstructions, and their corresponding accuracy(%)

CS Reconstruction	All	Yes/No	Number	Other
$\phi_B^T \phi_B x_B$***(MR = 0.25)***	48.92	70.61	33.13	36.58
ReconNet (MR = 0.25)	49.85	70.50	33.32	38.52
Oracle VQA v2.0				
LSTM + VGG	54.22	73.46	35.18	41.83
Question Only	44.26	67.01	31.55	27.37

CS Reconstructions



Conclusion

- VQA can achieve near-equivalent performance to natural images when using advanced compressive sensing reconstruction techniques such as ReconNet
- Using direct inference approaches, we report reduced processing time and network parameters over approaches that need full reconstruction. Of course, using a full-reconstruction approach results in the best performance

[1] M. Wakin, J. Laska, M. Duarte, D. Baron, S. Sarvotham, D. Takhar, K. Kelly, R. Baraniuk, "An architecture for compressive imaging", in *Proceedings of International Conference on Image Processing (ICIP),* 2006.

[2] K. Kulkarni, S. Lohit, P. Turaga, R. Kerviche, and A. Ashok, "ReconNet: Non-iterative reconstruction of images from compressively sensed random measurements," in *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR),* 2016.

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions" in *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR),* 2015.

[4] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "VQA: Visual question answering" in *Proceedings of International Conference on Computer Vision (ICCV),* 2015.

[5] Y. Goyal, T. Khot, D. Summers-Stay, D.Batra, and Devi Parikh, "Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering," in *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR), 2017*

Scan me