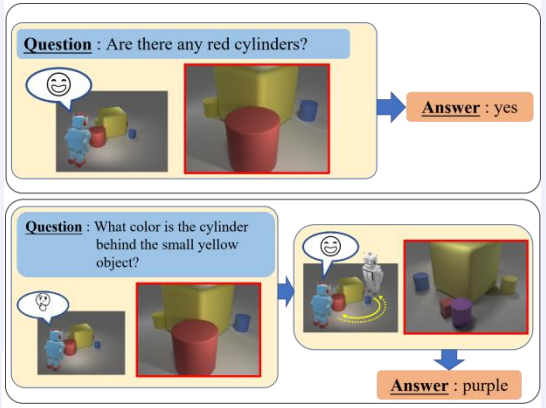
Multi-View Visual Question Answering with Active Viewpoint Selection

Yue Qiu [1,2], Yutaka Satoh [2,1], Ryota Suzuki [2], Kenji Iwata [2], Hirokatsu Kataoka [2]
[1]University of Tsukuba, [2]National Institute of Advanced Industrial Science and Technology

Motivation

- Visual Question Answering:
  - A vision and language multi-modal task that aims at answering a given question regarding the content of a provided image.
- Conventional single view VQA:
  - less ability to recognize geometrical information, so that they tend to fail to count or decide spatial relationship.
  - less ability to determine blind space for working in highly-occluded real-world environments.

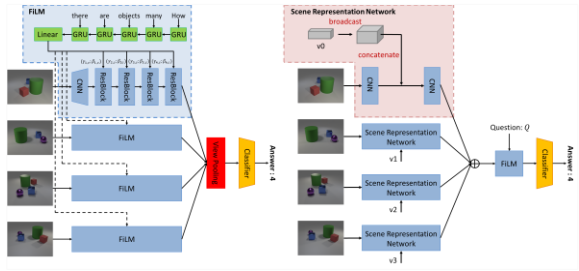
Introduction



- A multi-view VQA framework with viewpoint selection
- A DNN archtecture incorporating VQA, Scene Representation and Viewpoint selection modules
- Results:
  - Keeps performance against VQA using all views
  - Reduces trivial observation largely
  - Applicable on both CG and Real Images settings
  - Run time : 0.035 sec / VQA sample on average

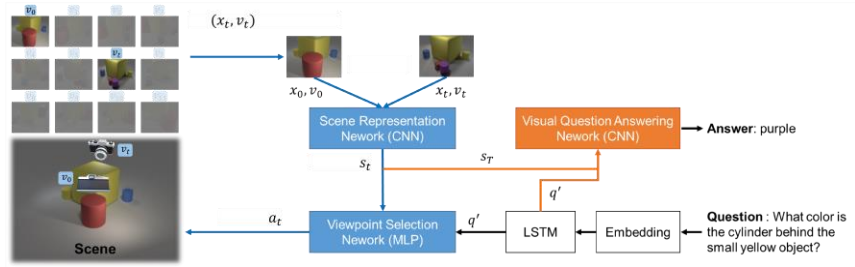
Approach

Approach 1 : View pooling + FiLM (VQA Backbone)
Approach 2 : Scene Representation Network + FiLM



- View pooling : max/average pooling for integrating multi-view image features.
- Scene Representation Network : CVAE-based structure for representing 3D scene and rendering images from query viewpoint.
- FiLM : Feature-wise Linear Modulation.

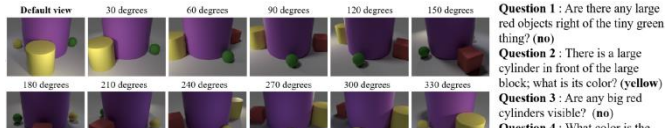
Approach 3: Scene Representation Network + Viewpoint Selection Network + FiLM

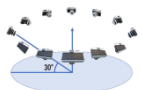


- Viewpoint Selection Network （Deep Q-learning Network - based）:
  - ✓ Input: Observed scene representation, question feature
  - ✓ Output: Next observation viewpoint.
  - ✓ Reward: Based on VQA accuracy and trajectory length.

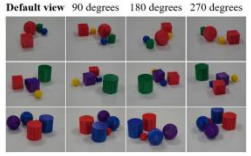
Dataset

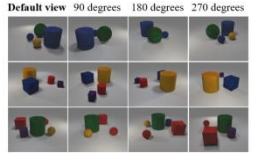
- CG Dataset:
  - ✓ Image generation: Place objects on blender scene.Photograph from multiple observation viewpoint.
  - ✓ Question generation: Generated automatically from function programs based on scene information.
- Real Dataset:
  - ✓ Image generation: Real photograph.
  - ✓ Question generation : Generated from function program.





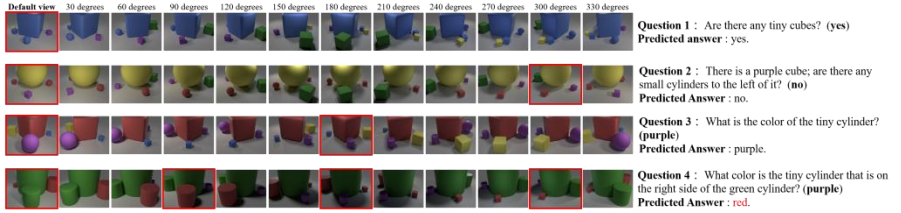
Virtual camera arragement





Experiments

Example results of SRN FiLM VS on Multi-view-CLEVR 12views CG dataset



Results on Multi-view-CLEVR_12views_CG

Methods	Overall accuracy	Spatial-related		Non-spatial		Average used viewpoints
		Exist	Query color	Exist	Query color	
SRN_FiLM	**97.37%**	94.20%	98.20%	99.03%	98.11%	12
SRN_FiLM_VS	97.11%	95.27%	96.90%	99.03%	97.25%	**2.98**

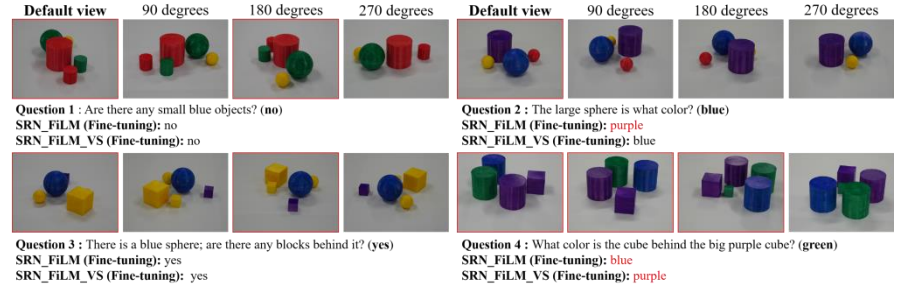
Effect of viewpoint selection

Methods (viewpoint numbers)	Overall accuracy
SRN_FiLM_Random (**3 views**)	81.39%
SRN_FiLM_Equal (**3 views**)	82.90%
SRN_FiLM_VS (**2.98 views**)	**97.11%**

Results on Multi-view-CLEVR_4views_CG

Methods	Accuracy	Used Viewpoints
SRN_FiLM	**97.67%**	4
SRN_FiLM_VS	97.64%	**2.02**

Example results on Multi-view-CLEVR 12views Real dataset



Results on Multi-view-CLEVR_4views_Real

Methods	Fine-tuning		Accuracy
	SRN	FiLM	
SRN_FiLM	-	-	67.88%
SRN_FiLM	-	✓	76.14%
SRN_FiLM	✓	-	77.30%
SRN_FiLM	✓	✓	82.62%
SRN_FiLM_VS	-	-	66.82%
SRN_FiLM_VS	-	✓	79.99%
SRN_FiLM_VS	✓	-	91.56%
SRN_FiLM_VS	✓	✓	**94.01%**

References

[1] Yue Qiu, Yutaka Satoh, Ryota Suzuki, Kenji Iwata and Hirokatsu Kataoka, H. Multi-View Visual Question Answering with Active Viewpoint Selection. Sensors 2020, 20, 2281.
[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In Proceedings of the IEEE international conference on computer vision (ICCV), 2015.
[3] Yue Qiu, Yutaka Satoh, Ryota Suzuki, and Hirokatsu Kataoka. Incorporating 3d information into visual question answering. In International Conference on 3D Vision (3DV), 2019.
[4] Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S Morcos, Marta Garnelo, Avraham Ruderman, Andrei A Rusu, Ivo Danihelka, Karol Gregor, et al.
Neural scene representation and rendering.Science, 2018.
[5] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In Thirty-Second AAAI Conference on Artificial Intelligence (AAAI), 2018.
[6] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. In Advances in Neural Information Processing Systems (NIPS), 2013.
[7] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2017.