Visual Question Answering for Relational Reasoning

Srikanth Grandhe, Shruti Gullapuram, Srideepika Jayaraman, Anish Pimpley

UMass CICS, Microsoft Research Montreal | Mentors : Samira Kahou, Adam Atkinson, Adam Trischler

Objective

To develop a model capable of **Visual and Relational Reasoning on the Figure QA dataset**

Task and Background

The VQA task deals with answering natural language queries about objects in images. Relational reasoning in VQA focuses specifically on queries involving relationships between objects in images. Performance on such a task hinges primarily on how **interaction between image and language representations** is modeled. Queries in below mentioned datasets concern **spatial relations, color, texture, shape and statistical properties** of objects in an image.

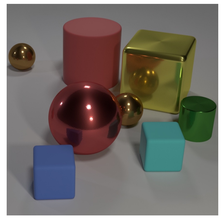
Sample Questions:

Are there an equal number of large things and metal spheres?

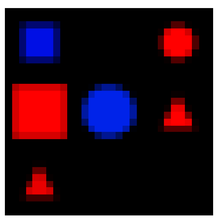
How many objects are either small cylinders or red things?

Dataset	#Img	#Qn	#Cat	Desc.
FigureQA	120K	1.5M	2	Plots & Charts
CLEVR	100K	850M	28	3D shapes
SHAPES	15K	300	2	2D shapes

Table 1: Dataset Overview

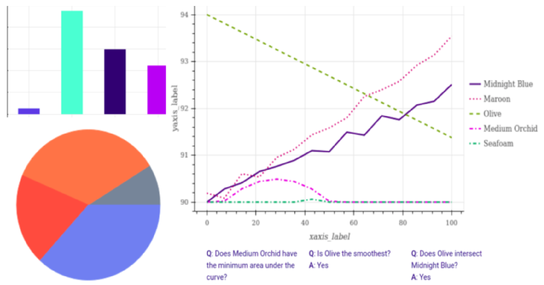


(a) CLEVR



(b) SHAPES

(c) sort-of-clevr



Approaches

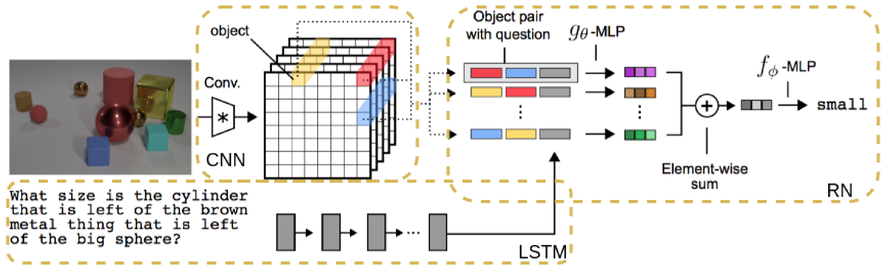
Relation Nets: Exhaustive Pair-wise Object Comparison

- **-Group Attention** to Reduce Object pairs
- **-Convolutional Attention** on image features
- **-Conditional Batch Norm** on intermediate features

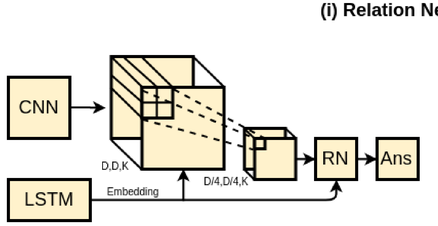
FiLM: Feature-wise Linear Modulation for Conditioning

- **-Stepwise generation** of CBN layers

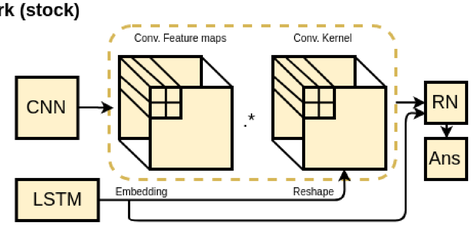
Model Architecture



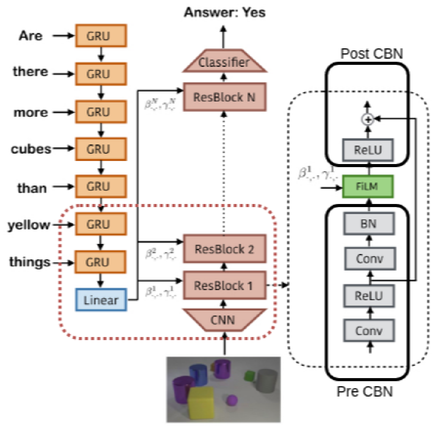
(i) Relation Network (stock)



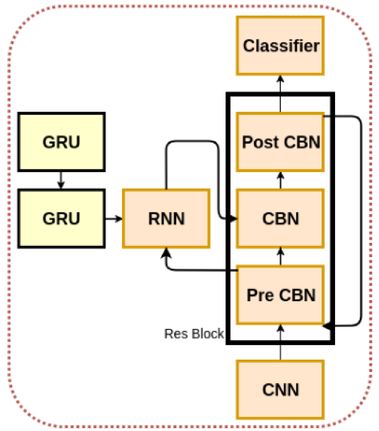
(ii) Group Attention



(iii) Convolutional Kernel

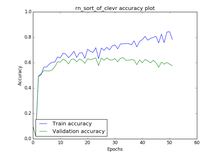


(iv) FiLM (stock)

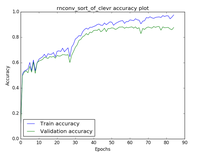


(v) Step-wise CBN generation

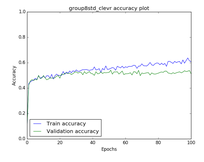
Experiments & Results



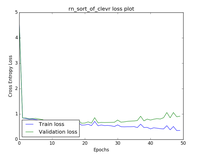
(e) RN accuracy on SOC



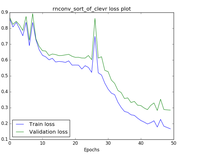
(f) Convolutional Attention accuracy on SOC



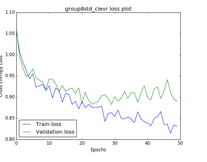
(g) Group Attention accuracy on CLEVR



(h) RN loss for SOC



(i) Convolutional attention loss for SOC



(j) Group Attention loss for CLEVR

Models	SOC	CLEVR	FigureQA
CNN + LSTM	N/A	51.2	50.5
Relation Network	64.9	51.9	53.3
Group Attention	79.1	**54.1**	**56.3**
Convolutional Attention	**87.8**	53.3	55.2
CBN in RN	66.4	52.4	54.9

Table 2: Performance of models on different datasets

Future Work

- Evaluate on complete FigureQA dataset
- Reduce pairwise object comparisons in Relation Networks
- Stacked co-attention model
- Tournament structured RN

References

- E. Perez et al. Film: Visual reasoning with a general conditioning layer.
- A. Santoro et al. A simple neural network module for relational reasoning.
- S.E.Kahou et al.,FigureQA: An annotated figure dataset for visual reasoning.
- J. Johnson et al. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning.