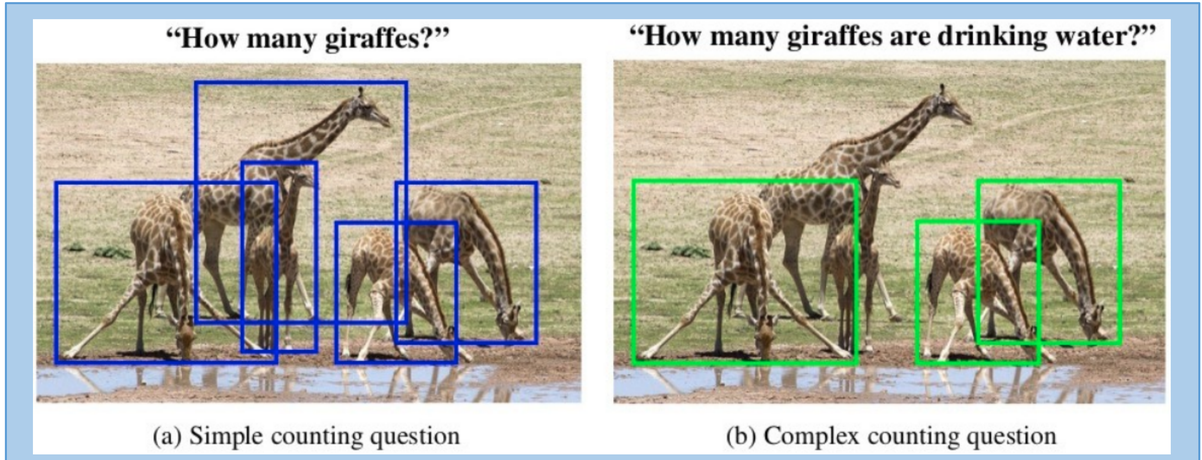
TallyQA: Answering Complex Counting Questions

Manoj Acharya, Kushal Kafle, and Christopher Kanan
Rochester Institute of Technology, Rochester NY
manoj@mail.rit.edu, kk6055@rit.edu, kanan@rit.edu

Overview



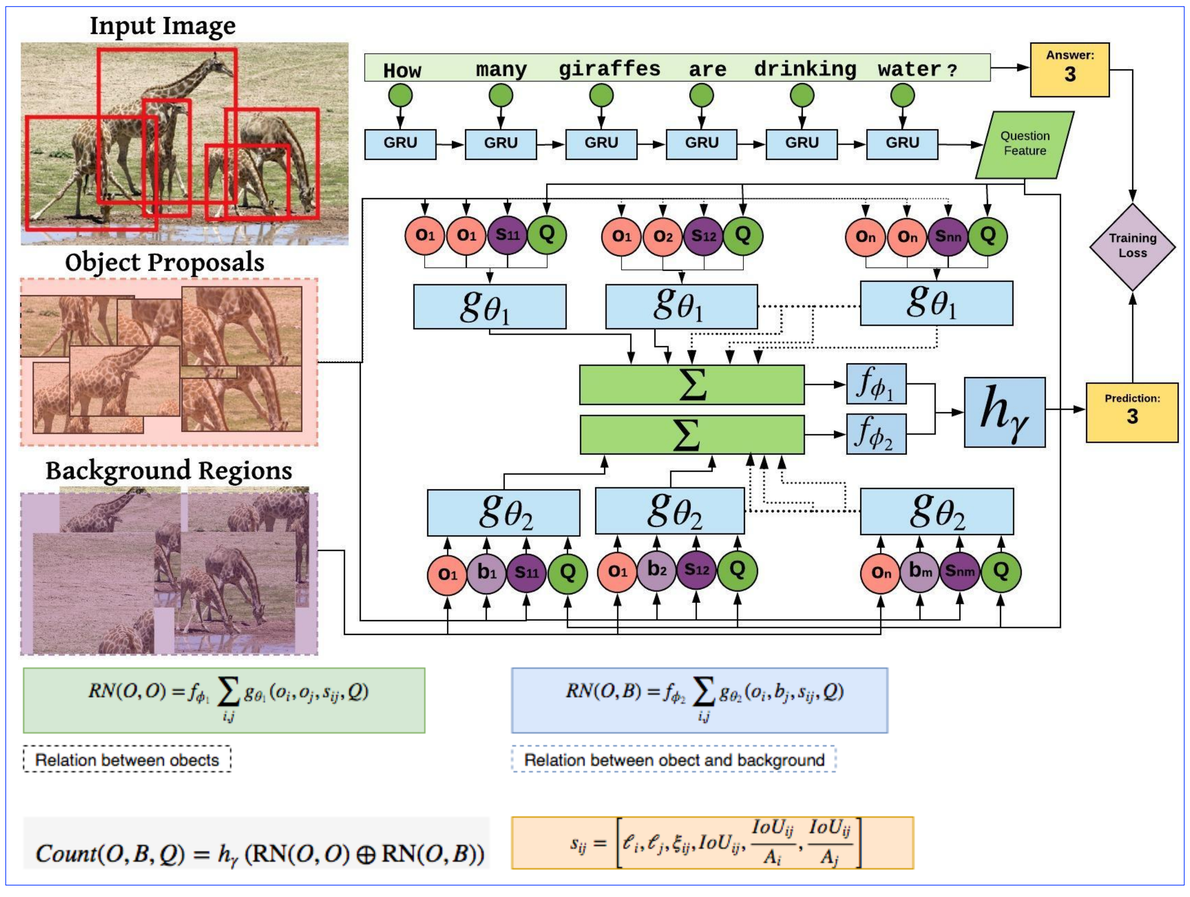
- State-of-the-art Visual Question Answering (VQA) systems now rival humans on many kinds of questions, but counting performance is comparatively poor: 72% overall on VQA2, but only 51% for counting.
- Not only is performance poor, most counting questions in existing datasets are simple - they can be solved using only object detection. We address this problem.
- Our major contributions:
  - Created TallyQA, which is almost twice as large as the next biggest counting dataset. It has both simple and complex questions.
  - Created the novel RCN algorithm for open-ended counting.
  - RCN achieves state-of-the-art results on TallyQA and earlier datasets.

TallyQA

TallyQA is the **largest open-ended counting dataset by a factor of two**. It distinguishes between simple and complex counting questions.It has:

- 288K Question-Answer pairs
- 165K Images
- 19.5K complex QA pairs collected from human annotators using AMT

Relational Network for Counting (RCN)



RCN Features

- 26x faster than naive RN networks
- RCN models the interaction between objects and backgrounds, which helps it understand scene context better
- 6.3% improvement on positional reasoning questions by using background information

Results

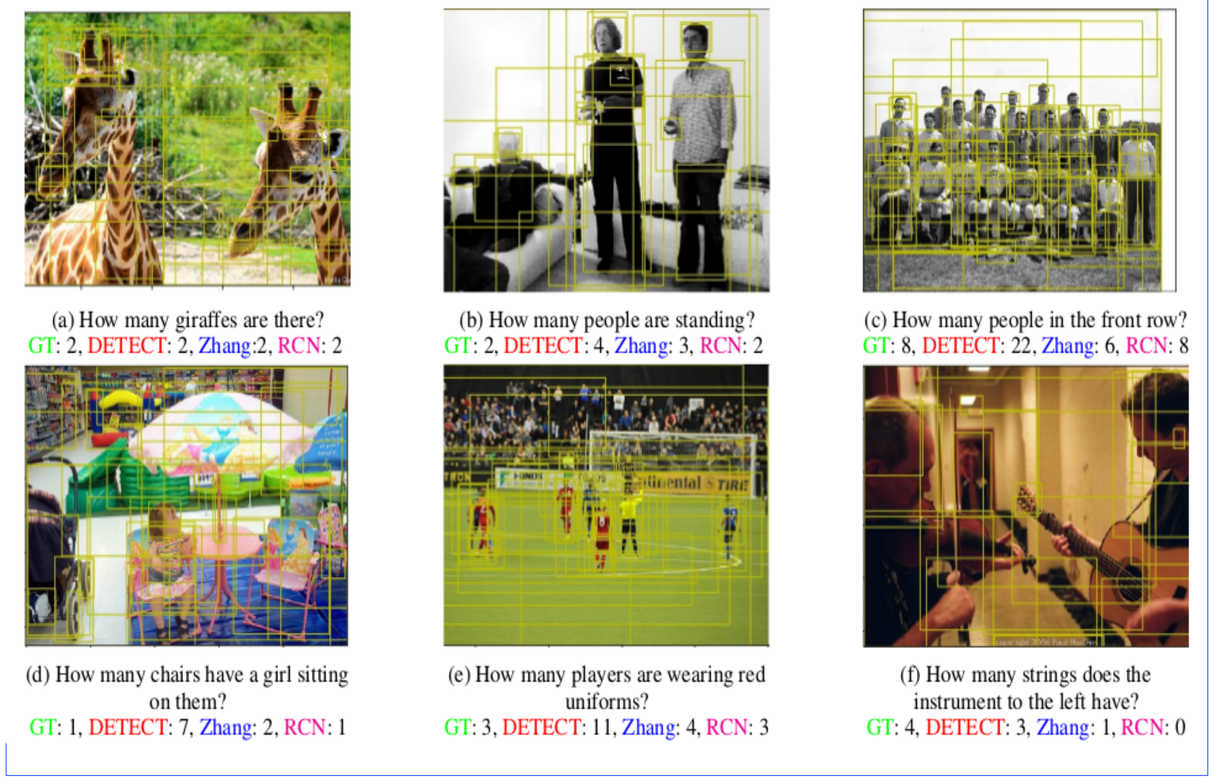
	HowMany-QA		TallyQA Test-Simple		TallyQA Test-Complex	
	ACC	RMSE	ACC	RMSE	ACC	RMSE
Guess-1	33.8	3.74	53.5	1.78	43.9	1.57
Guess-2	32.1	3.34	24.5	1.56	15.9	1.69
Q-Only	37.1	3.51	44.6	1.74	39.1	1.75
I-Only	37.3	3.49	46.1	1.71	26.4	1.69
Q+I	40.5	3.17	54.7	1.44	48.8	1.57
DETECT	43.3	3.66	50.6	2.08	15.0	4.52
MUTAN	45.5	2.93	56.5	1.51	49.1	1.59
Zhang *et al.*	54.7	2.59	70.5	1.15	50.9	1.58
IRLC	56.1	2.45	–	–	–	–
RCN (Ours)	**60.0**	**2.39**	**71.8**	**1.13**	**56.2**	**1.43**

- MUTAN: State-of-the-art VQA model in 2016
- IRLC: State-of-the-art counting model on HowMany-QA dataset
- Zhang *et al.*: State-of-the-art counting model for VQA2 dataset

Conclusion and Future Work

- Open Ended Counting is still far from solved
- Complex questions provide insight about VQA models
- Need to make models interpretable, evaluate compositional reasoning, and analyze their ability to count values at test time that were not seen during training.

Example Predictions



(a) How many giraffes are there?
GT: 2, DETECT: 2, Zhang:2, RCN: 2

(b) How many people are standing?
GT: 2, DETECT: 4, Zhang: 3, RCN: 2

(c) How many people in the front row?
GT: 8, DETECT: 22, Zhang: 6, RCN: 8

(d) How many chairs have a girl sitting on them?
GT: 1, DETECT: 7, Zhang: 2, RCN: 1

(e) How many players are wearing red uniforms?
GT: 3, DETECT: 11, Zhang: 4, RCN: 3

(f) How many strings does the instrument to the left have?
GT: 4, DETECT: 3, Zhang: 1, RCN: 0