An Empirical Evaluation of Visual Question Answering for Novel Objects

Santhosh K. Ramakrishnan[1,2], Ambar Pal[1], Gaurav Sharma[1], Anurag Mittal[2]
[1]IIT Kanpur, [2]IIT Madras

IEEE 2017 Conference on Computer Vision and Pattern Recognition
CVPR July 21-26 2017 HONOLULU

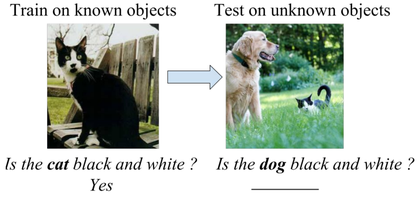
Motivation

Existing work on VQA focuses on datasets where the **train and test objects overlap significantly**. In real world scenarios, there are a large number of objects for which (image, question, answer) data is not available. This work is the first to **generalize current VQA systems** to objects not queried about in training data.

Problem Description

Given
- VQA dataset with **test objects not present in train set (novel)** .
- Unpaired external text and image data,

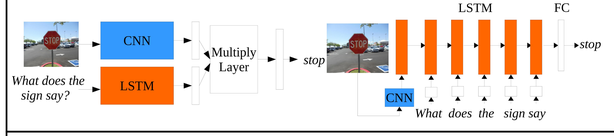
Answer test questions containing these novel objects.



Datasets

- VQA dataset
- BookCorpus, Wikipedia dump (external text)
- ImageNet (external images)

Architecture 1 **Architecture 2**



Novel split of VQA dataset

- List out nouns from questions and answers in train set
- Cluster nouns based on their question type statistics
- Separate out 20% nouns from each cluster as test nouns
- Assign all questions containing test nouns to test set and the rest to train set

Split	# Questions Train	# Questions Test	# Objects Train	# Objects Test	% Novel
Orig	215375	121509	3625	3330	4.6
Prop	224704	116323	2951	3027	26.8

- **~27%** test objects are novel in proposed split

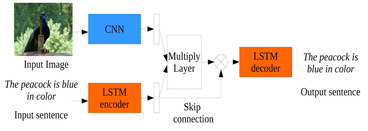
Baseline performance on Original vs Novel split

Split	Architecture 1 OpenEnded	Architecture 1 MultipleChoice	Architecture 2 OpenEnded	Architecture 2 MultipleChoice
Orig	54.23	59.30	48.75	54.94
Novel	39.38	46.54	34.97	42.83
Drop	**14.85**	**12.76**	**13.78**	**12.11**

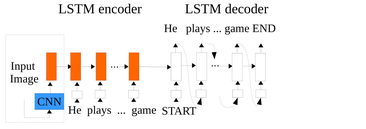
- Drop in overall accuracy of **~14%** in both the architectures !

Approach to induce novel objects

Seq2seq AutoEncoder Architectures



- Used for **weak paired** training in VQA **Arch 1**



- Used for **weak paired** training in VQA **Arch 2**
- Used for **text only** training in VQA **Archs 1 and 2**
- Pre-train image, text encoders on auxiliary data
- Finetune them on VQA

Text only induction
- Train text to text AutoEncoder (AE)
- Incorporate novel objects into vocabulary under 2 settings:
  - **Oracle:** novel words known textually
  - **General:** novel words semantically similar to known words

Weakly paired text + image induction
- Pair images of novel object with random text about it
- Train image + text to text AutoEncoder (AE)

Quantitative Results

- Feature - pre-trained VGGnet (VGG) or Google Inception (INC)
- Aux - Auxiliary data, none / text / weak paired (text + im)
- Vocab - train (only VQA train), oracle, general
- OEQ - Open Ended Questions, MCQ - Multiple Choice Questions

Text based novel object induction

Feature	Aux	Vocab	Architecture 1 (OEQ) Overall	Yes/No	Novel	Architecture 2 (OEQ) Overall	Yes/No	Novel
VGG	none	oracle	39.38	74.02	47.56	34.97	71.06	44.60
VGG	text	oracle	**40.44**	**76.52**	**48.95**	**37.68**	**75.06**	**46.93**
INC	none	oracle	40.27	73.95	48.03	37.66	73.69	46.50
INC	text	oracle	**41.19**	**75.93**	**49.23**	**38.53**	**75.39**	**47.55**

- Text data provides overall improvements of ~2 % in arch 1 and ~3% in arch 2
- Majority of improvement is in Yes/No and Novel question types

Weakly paired data based novel object induction

Feature	Aux	Vocab	Architecture 1 OEQ	Architecture 1 MCQ	Architecture 2 OEQ	Architecture 2 MCQ
VGG	text	oracle	40.44	**47.65**	37.68	45.12
VGG	text+im	oracle	**40.49**	47.38	**38.06**	**45.80**
VGG	text	gen(exp)	**40.76**	**47.82**	**38.00**	**45.96**
VGG	text+im	gen(exp)	40.34	47.36	37.92	45.58
INC	text	oracle	**41.19**	**47.87**	38.53	45.85
INC	text+im	oracle	40.73	47.23	**38.75**	**46.07**
INC	text	gen(exp)	**41.39**	**47.88**	37.99	**45.89**
INC	text+im	gen(exp)	40.42	46.87	**38.20**	45.65

- Provides marginal improvements in arch 2 over text only induction
- A noisy method of pairing the data

Need to incorporate novel words into vocabulary

Feature	Aux	Vocab	Architecture 1 OEQ	Architecture 1 MCQ	Architecture 2 OEQ	Architecture 2 MCQ
VGG	none	oracle	39.38	46.54	34.97	42.83
VGG	text	train	40.09	47.22	37.30	44.30
VGG	text	oracle	**40.44**	**47.65**	**37.68**	**45.12**
INC	none	oracle	40.27	46.47	37.66	44.59
INC	text	train	40.18	47.01	37.37	44.40
INC	text	oracle	**41.19**	**47.87**	**38.53**	**45.85**

- Necessary to incorporate novel objects into vocabulary
- Train words + external data can even lead to poorer performance

Additional observations
- Using Inception features over VGG features does not improve novel VQA performance
- Using pre-trained word vectors to expand vocabulary in general setting helps
- Improvement obtained on the better architecture (arch 1) is unfortunately lesser

Qualitative Results

P - Proposed, B - Baseline, GT - Ground Truth

Positive Examples

Negative Examples

Conclusions

- Challenging and real-world setting that needs to be addressed
- Significant drop in performance of two existing architectures in the new setting
- Proposed 2 methods for inducing novel objects
- Text based induction - effective; Weak pairing - not effective (noisy)
- External text data without novel object induction need to be effective

Contact: ee12b101@ee.iitm.ac.in (Santhosh) **Project Site: https://goo.gl/ELLb9z**