An Empirical Evaluation of Visual Question Answering for Novel Objects
Santhosh K. Ramakrishnan1,2, Ambar Pal1, Gaurav Sharma1, Anurag Mittal2
IIT Kanpur, 2IIT Madras

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

<table>
<thead>
<tr>
<th>Feature</th>
<th>Aux</th>
<th>Vocab</th>
<th>OEQ</th>
<th>MCQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGG</td>
<td>None</td>
<td>oracle</td>
<td>39.38</td>
<td>42.74</td>
</tr>
<tr>
<td>VGG</td>
<td>text</td>
<td>oracle</td>
<td>40.44</td>
<td>42.69</td>
</tr>
<tr>
<td>INC</td>
<td>text</td>
<td>oracle</td>
<td>41.19</td>
<td>45.78</td>
</tr>
<tr>
<td>INC</td>
<td>text</td>
<td>general</td>
<td>40.44</td>
<td>46.87</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Feature</th>
<th>Aux</th>
<th>Vocab</th>
<th>OEQ</th>
<th>MCQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGG</td>
<td>text</td>
<td>gen.</td>
<td>39.38</td>
<td>42.74</td>
</tr>
<tr>
<td>VGG</td>
<td>text</td>
<td>gen.</td>
<td>40.44</td>
<td>42.69</td>
</tr>
<tr>
<td>INC</td>
<td>text</td>
<td>gen.</td>
<td>41.19</td>
<td>45.78</td>
</tr>
<tr>
<td>INC</td>
<td>text</td>
<td>gen.</td>
<td>40.44</td>
<td>46.87</td>
</tr>
</tbody>
</table>

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

Weakly paired text + image induction

<table>
<thead>
<tr>
<th>Feature</th>
<th>Aux</th>
<th>Vocab</th>
<th>OEQ</th>
<th>MCQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGG</td>
<td>None</td>
<td>oracle</td>
<td>39.38</td>
<td>42.74</td>
</tr>
<tr>
<td>VGG</td>
<td>text</td>
<td>oracle</td>
<td>40.44</td>
<td>42.69</td>
</tr>
<tr>
<td>INC</td>
<td>text</td>
<td>oracle</td>
<td>41.19</td>
<td>45.78</td>
</tr>
<tr>
<td>INC</td>
<td>text</td>
<td>oracle</td>
<td>40.44</td>
<td>46.87</td>
</tr>
</tbody>
</table>

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

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.

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

Architecture 1
- LSTM encoder
- LSTM decoder

Architecture 2
- LSTM encoder
- LSTM decoder

Pre-train image, text encoders on auxiliary data
Fine-tune them on VQA

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

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

Architecture 1
- Used for weak paired training in VQA Arch 1
- Used for text only training in VQA Archs 1 and 2

Architecture 2
- Used for weak paired training in VQA Arch 2
- Used for text only training in VQA Arch 1 and 2

Text only induction
➢ Train text to text AutoEncoder (AE)
➢ Incorporate novel objects into vocabulary under 2 settings:
  o Oracle: novel words known textually
  o 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)

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

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

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