Being Negative but Constructively: Lessons Learnt from Creating Better Visual Question Answering Datasets
Wei-Lun (Harry) Chao*, Hexiang (Frank) Hu*, and Fei Sha
U. of Southern California

Highlights
- **Goal:** How to design good visual QA dataset?
- **Observation:** On existing multiple-choice (MC) datasets, models can ignore information while still doing well.
- **Insight:** The design of negative answers (decoys) significantly affects the learning behavior.
- **Contributions:** Propose principles and automatic procedures to generate decoys, remedying two popular datasets (VQA, Visual7W) as well as creating a new one based on the Visual Genome (VG) projects.
- **Link:** [http://www.teds.usc.edu/website_vqa/](http://www.teds.usc.edu/website_vqa/)

Introduction
- Multiple-choice Visual Question Answering (QA): Given an image (I), a question (Q), and a candidate answers set (A)—a target (T)—k decoys (D)—a machine needs to select the correct one.
- **Goal:** comprehend and reason with visual + language info.

Q: What vehicle is pictured?
A: A car. (0.21)
A bus. (0.62)
A cab. (0.50)
A train. (0.73)

How to design decoys is rarely discussed: random, high frequency, or human generated ones by looking at Q and T

Analysis
- **Dataset:** Visual7W (each IQA triplet has 4 candidates (C))
- **VQA model:**
  - MLP to predict the score of each IQC triplet
  - Features: CNN for I, WORD2VEC for Q and C, by concatenation

Principles and automatic procedures
- **Principles**
  - **Neutrality** (remove incidental statistics)
  - **QoU** (question only unsolvable)
  - **IoU** (image only unsolvable)

Automatic procedures
- **Requirements:** (1) IQT triplets are provided.
  - (2) I with multiple QT pairs
- **QoU-decoys:** targets of similar Q'
- **IoU-decoys:** targets of Q' of the same I
- **Resolve ambiguity:** (1) string matching
  - (2) Wu-Palmer scores

Diagnosis
- **IoU decoys**
  - Overcast. (0.55)
  - Daytime. (0.49)
  - A building. (0.48)
  - A train. (0.54)

Experiments
- **Dataset**
  - # images
  - # triplets
  - # Orig. D

<table>
<thead>
<tr>
<th>Method</th>
<th>Orig</th>
<th>IoU</th>
<th>QoU</th>
<th>IoU+QoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML-P &amp; ML-P</td>
<td>92.9</td>
<td>27.0</td>
<td>34.1</td>
<td>15.7</td>
</tr>
<tr>
<td>ML-P/A</td>
<td>62.4</td>
<td>73.3</td>
<td>55.0</td>
<td>23.6</td>
</tr>
<tr>
<td>ML-P/QA</td>
<td>58.2</td>
<td>84.1</td>
<td>47.3</td>
<td>37.8</td>
</tr>
<tr>
<td>ML-P/VG</td>
<td>65.7</td>
<td>84.1</td>
<td>57.6</td>
<td>52.0</td>
</tr>
<tr>
<td>Human</td>
<td>88.4</td>
<td>-</td>
<td>84.1</td>
<td>-</td>
</tr>
<tr>
<td>Random</td>
<td>25.0</td>
<td>25.0</td>
<td>25.0</td>
<td>14.3</td>
</tr>
</tbody>
</table>

#C as T (0.51)
#C as T+(#C as D)/K

- **Machines need to use all three information** (i.e., I, Q, A) to perform well.