Natural Language for Visual Reasoning for Real:

NLVR$^2$

Alane Suhr*, Stephanie Zhou*, Iris Zhang, Huajun Bai, Yoav Artzi

*Equal contribution
Reasoning about Language and Vision

What is the dog carrying?
(VQA, Agrawal et al. 2015)

Are there an equal number of large things and metal spheres?
(CLEVR, Johnson et al. 2017)

there are exactly three squares not touching any edge
(NLVR, Suhr et al. 2017)
Reasoning about Language and Vision

**Our goal:** Large corpus of natural language paired with photographs focusing on a diverse set of linguistic phenomena

---

*What is the dog carrying?* (VQA, Agrawal et al. 2015)

*Are there an equal number of large things and metal spheres?* (CLEVR, Johnson et al. 2017)

*there are exactly three squares not touching any edge* (NLVR, Suhr et al. 2017)
Natural Language for Visual Reasoning for Real (NLVR²)

Task: Determine whether the sentence is true or false about the pair of images.

All dogs are corgis with upright ears, and one image contains at least twice as many real corgis as the other image.
Natural Language for Visual Reasoning for Real (NLVR$^2$)

Task: Determine whether the sentence is true or false about the pair of images.

All dogs are corgis with upright ears, and one image contains at least twice as many real corgis as the other image.

TRUE
Outline

1. Task
2. Data collection
3. Analysis
4. Results
Task

All dogs are corgis with upright ears, and one image contains at least twice as many real corgis as the other image.

The left and right image contains the same number of white framed door window panes on at least one door way.

TRUE

FALSE
Data Collection

Our goal: Natural language descriptions of photographs and truth judgments

1. Background: NLVR data collection
2. Image collection
3. Sentence writing
4. Validation
Background:
NLVR Image Generation
Background: NLVR Image Generation

- Randomly generate a single image
Background:
NLVR Image Generation

- Randomly generate a single image
- Randomly generate another image
Background: NLVR Image Generation

- Randomly generate a single image
- Randomly generate another image
Background: NLVR Image Generation

- Randomly generate a single image
- Randomly generate another image
- Generate a third image, using objects from top image
NLVR Image Generation

- Randomly generate a single image
- Randomly generate another image
- Generate a third image, using objects from top image
- Generate a fourth image similarly
There is a box with 3 items of all 3 different colors.
There is a box with 3 items of all 3 different colors.

There is a box with 3 items of all 3 different colors.

There is a box with 3 items of all 3 different colors.

There is a box with 3 items of all 3 different colors.
Key Data Collection Challenges

• In NLVR, can control image generation to enable complex reasoning

• Can’t generate real images

• How can we use real images but ensure images are visually complex?
  • Queries that elicit complex images
  • Similar Image tools
1. **Pick 124 synsets from ImageNet**

   Chose synsets that would often appear multiple times in one image: e.g., acorn >> sump pump

   - Allows use of ImageNet models and tools
   - Allows for weak annotation of image content
1. **Pick 124 synsets from ImageNet**
   Chose synsets that would often appear multiple times in one image: e.g., acorn >> sump pump
   - Allows use of ImageNet models and tools
   - Allows for weak annotation of image content
1. **Pick 124 synsets from ImageNet**
   Chose synsets that would often appear multiple times in one image: e.g., acorn >> sump pump

2. **Generate and execute search queries**
   Combined synset names with numerical phrases, hypernyms, and similar words

   ![two acorns](image-url)
Image Collection

1. Pick 124 synsets from ImageNet
   Chose synsets that would often appear multiple times in one image: e.g., acorn >> sump pump

2. Generate and execute search queries
   Combined synset names with numerical phrases, hypernyms, and similar words
1. **Pick 124 synsets from ImageNet**
   Chose synsets that would often appear multiple times in one image: e.g., acorn >> sump pump

2. **Generate and execute search queries**
   Combined synset names with numerical phrases, hypernyms, and similar words
Image Collection

3. Remove low-quality images
   Don’t contain synset, drawings, inappropriate content
Image Collection

3. Remove low-quality images
   Don’t contain synset, drawings, inappropriate content
3. **Remove low-quality images**
   Don’t contain synset, drawings, inappropriate content

4. **Construct sets of eight images**
   Each set must contain at least three *interesting* images (e.g., multiple objects)
4. **Construct sets of eight images**
   Each set must contain at least three *interesting* images (e.g., multiple objects)
4. **Construct sets of eight images**
Each set must contain at least three *interesting* images (e.g., multiple objects)
4. **Construct sets of eight images**
Each set must contain at least three *interesting* images (e.g., multiple objects)
Sentence Writing

5. Display a set of randomly paired images
Sentence Writing

5. Display a set of randomly paired images

6. Ask workers to select two pairs
Sentence Writing

5. Display a set of randomly paired images

6. Ask workers to select two pairs

7. Workers write a sentence true about the selected pairs, but false about the others

One image shows exactly two brown acorns in back-to-back caps on green foliage.
8. Show each image/sentence pair to another work and ask them to label it.

One image shows exactly two brown acorns in back-to-back caps on green foliage.

True  □
False □
Validation

8. Show each image/sentence pair to another work and ask them to label it.

One image shows exactly two brown acorns in back-to-back caps on green foliage.

True ☑️
False ✗
Data Collection and Corpus Statistics

• 107,296 total examples
  • 29,680 unique sentences
  • 127,506 unique images
  • 80% train, 20% evenly split among dev and two test sets

• Agreement: near perfect ($\alpha = 0.912$, $\kappa = 0.889$)

• Total cost: $19,282.99

• Average sentence length: 14.8 tokens

• Vocabulary size: ~7,500 word types
## Related Corpora

<table>
<thead>
<tr>
<th>Task</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MSCOCO</strong></td>
<td><strong>Caption generation</strong></td>
</tr>
<tr>
<td>(Chen et al 2015)</td>
<td>A small herd of cows in a large grassy field.</td>
</tr>
<tr>
<td><strong>VQA</strong></td>
<td><strong>Question answering</strong></td>
</tr>
<tr>
<td>(Agrawal et al 2015)</td>
<td>What is the dog carrying?</td>
</tr>
<tr>
<td><strong>CLEVR</strong></td>
<td><strong>Question answering</strong></td>
</tr>
<tr>
<td>(Johnson et al 2017a)</td>
<td>How many objects are either small cylinders or red things?</td>
</tr>
<tr>
<td><strong>CLEVR-Humans</strong></td>
<td><strong>Question answering</strong></td>
</tr>
<tr>
<td>(Johnson et al 2017b)</td>
<td>How many objects are not purple and not metallic?</td>
</tr>
<tr>
<td><strong>NLVR</strong></td>
<td><strong>Binary classification</strong></td>
</tr>
<tr>
<td>(Suhr et al 2017)</td>
<td>there are exactly three blue objects not touching any edge</td>
</tr>
<tr>
<td><strong>NLVR(^2)</strong></td>
<td><strong>Binary classification</strong></td>
</tr>
<tr>
<td></td>
<td>All dogs are corgis with upright ears, and one image contains at least twice as many real corgis as the other image.</td>
</tr>
</tbody>
</table>
## Related Corpora

<table>
<thead>
<tr>
<th>Corpo</th>
<th>Task</th>
<th>Real images?</th>
<th>Natural language?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MSCOCO</strong> (Chen et al 2015)</td>
<td>Caption generation</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td><strong>VQA</strong> (Agrawal et al 2015)</td>
<td>Question answering</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td><strong>CLEVR</strong> (Johnson et al 2017a)</td>
<td>Question answering</td>
<td>✗️</td>
<td>✗️</td>
</tr>
<tr>
<td><strong>CLEVR-Humans</strong> (Johnson et al 2017b)</td>
<td>Question answering</td>
<td>✗️</td>
<td>✔️</td>
</tr>
<tr>
<td><strong>NLVR</strong> (Suhr et al 2017)</td>
<td>Binary classification</td>
<td>✗️</td>
<td>✔️</td>
</tr>
<tr>
<td><strong>NLVR²</strong></td>
<td>Binary classification</td>
<td>✔️</td>
<td>✔️</td>
</tr>
</tbody>
</table>
Sentence Lengths

- NLVR\textsuperscript{2} (ours)
- NLVR
- MSCOCO
- VQA real images
- VQA abstract images
- CLEVR-Humans
- CLEVR

![Graph showing sentence lengths for different datasets](image-url)
Linguistic Analysis

- Hard cardinality: NLVR² 75, NLVR 30, VQA (real) 15
- Soft cardinality: NLVR² 30, NLVR 20, VQA (real) 15
- Coordination: NLVR² 30, NLVR 30, VQA (real) 15
- Negation: NLVR² 15, NLVR 15, VQA (real) 15
- Universal quantifiers: NLVR² 30, NLVR 20, VQA (real) 15
- Coreference: NLVR² 30, NLVR 30, VQA (real) 15
- Presupposition: NLVR² 80, NLVR 80, VQA (real) 80
- Spatial relations: NLVR² 70, NLVR 70, VQA (real) 70
- Comparisons: NLVR² 10, NLVR 10, VQA (real) 10
There are **two**, and only **two**, people.

There are no more than **eight** bottles in total.
Each image contains just one bird, and the wires of a cage are behind the bird in one image.
A mitten is being worn in one image and the mittens are not being worn in the other image.
Both images shows a silver pail being used as a flower vase.
the left image has 4 balloons of all different colors
Baselines

Accuracy on unreleased test set

Majority class: 51.4

Text only (RNN): 51.4

Image only (CNN): 51.9

CNN+RNN: 53.2

Object Detection MaxEnt: 53.5
SOTA Visual Reasoning

Accuracy on unreleased test set

<table>
<thead>
<tr>
<th>Model</th>
<th>Majority class</th>
<th>N2NMN (Hu et al. 2017)</th>
<th>FiLM (Perez et al. 2017)</th>
<th>MAC-Network (Hudson et al. 2018)</th>
<th>Object Detection MaxEnt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>51.4</td>
<td>51.5</td>
<td>53.0</td>
<td>51.2</td>
<td>53.5</td>
</tr>
<tr>
<td>CLEVR Accuracy</td>
<td></td>
<td>83.7%</td>
<td>97.7%</td>
<td>98.9%</td>
<td></td>
</tr>
</tbody>
</table>
Paper, Data, Code

• Paper, data, and code will be released soon

• Will integrate with ParlAI

• Thank you for attending

• Thanks to ParlAI for funding!