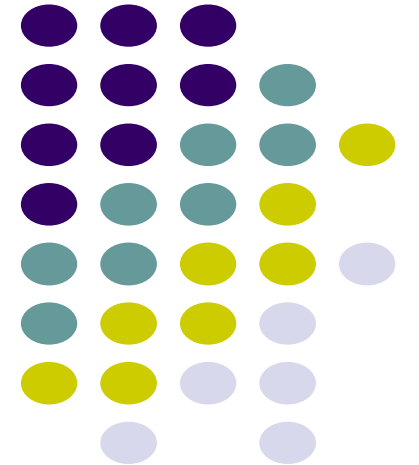


# A Corpus of Natural Language for Visual Reasoning

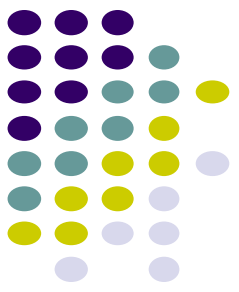
**Alane Suhr<sup>†</sup>, Mike Lewis<sup>‡</sup>, James Yeh<sup>†</sup>, and Yoav Artzi<sup>†</sup>**

<sup>†</sup> Dept. of Computer Science and Cornell Tech, Cornell University, New York, NY 10044  
`{suhr, yoav}@cs.cornell.edu, jamesclyeh@gmail.com`

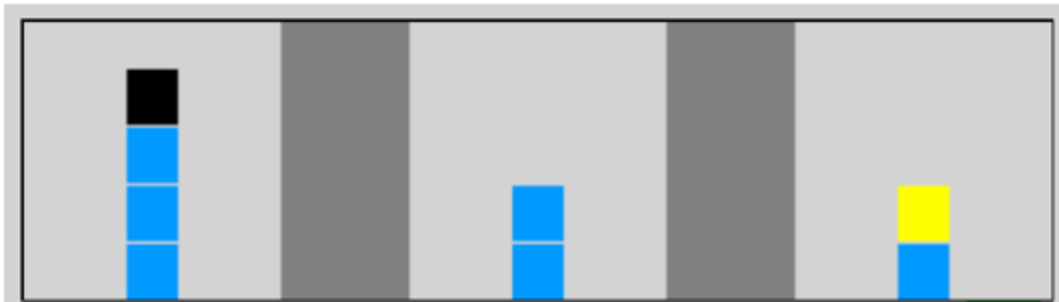
<sup>‡</sup> Facebook AI Research, Menlo Park, CA 94025  
`mikelewis@fb.com`



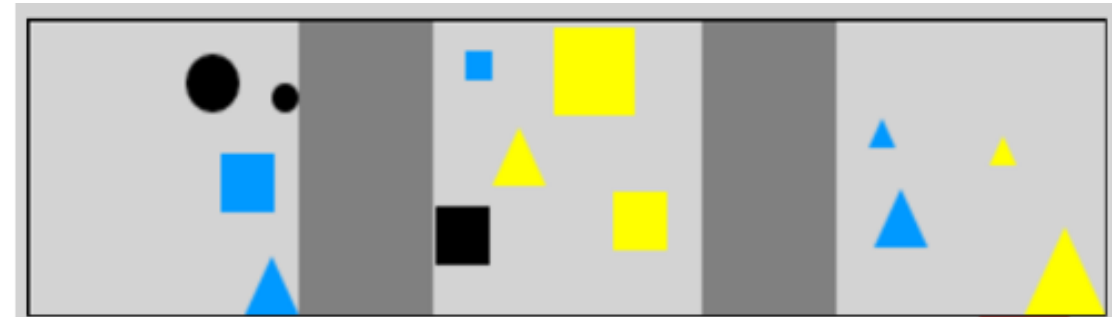
# Cornell Natural Language Visual Reasoning Dataset (NLVR)



- **Task:** Given a sentence-image pair, determine if a sentence is *true* or *false* about the image.



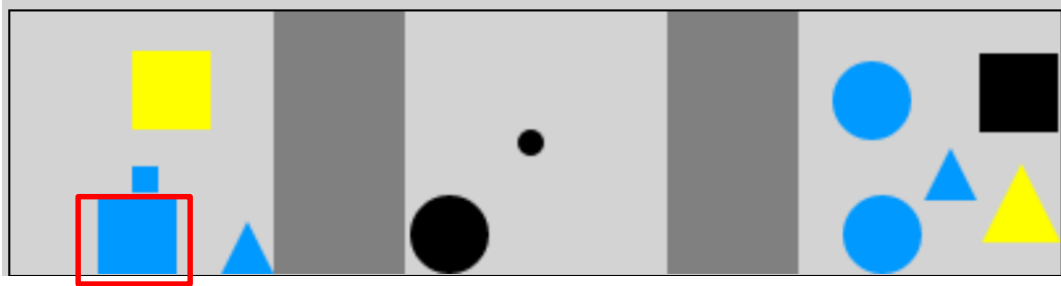
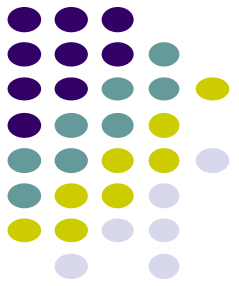
There is exactly one tower with a black block at the top ✓



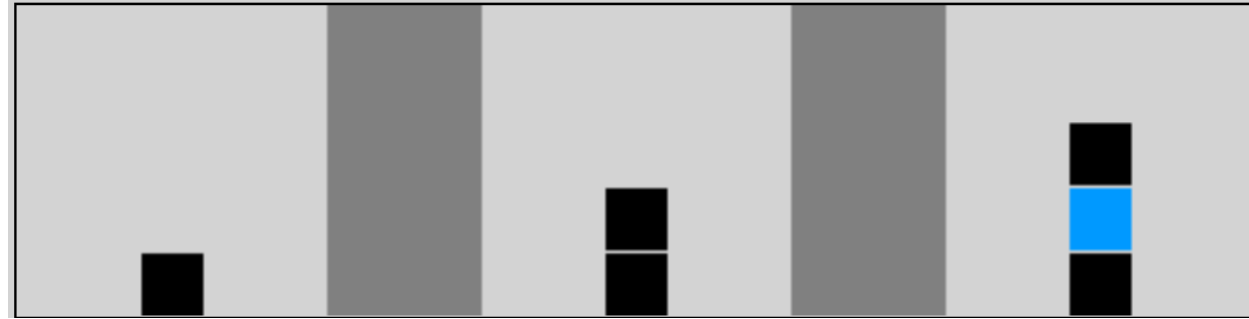
One of the grey boxes has exactly six objects ✗

- Requires reasoning about sets of **objects**, **quantities**, **colors** and **spatial relationships**
- **Applications:** Instructing assembly-line robots to manipulate objects in cluttered environments

# More Examples



There is blue square touching the bottom of a box ✓

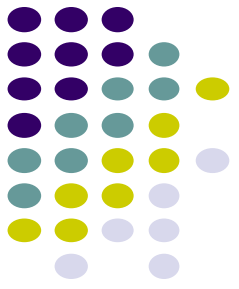


There is at least two towers with the same height ✗

## Goal of the paper:

- Describing the process of creating the dataset for this new task
- Reporting the results for several simple models trained on the dataset in order to show the complexity of the data

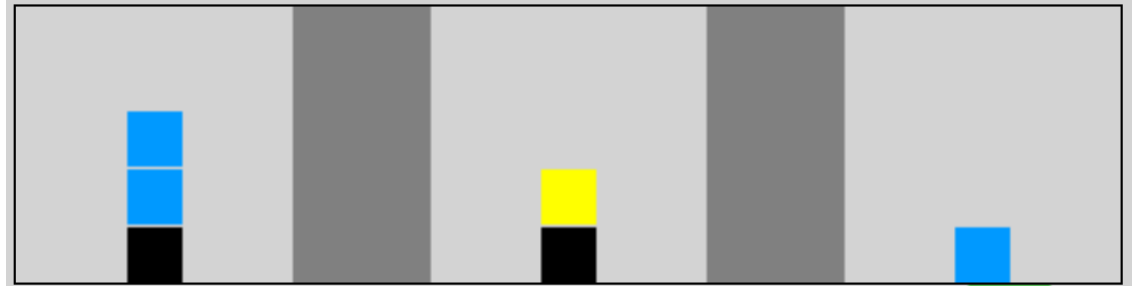
# Dataset Preparation



Generation of Structured Representation of each object in an image

```
{"type": "square", "color": "black", "x_loc": 40, "y_loc": 80, "size": "20"},  
{"type": "square", "color": "blue", ... },  
{ ..... }
```

Automatic Image Generation



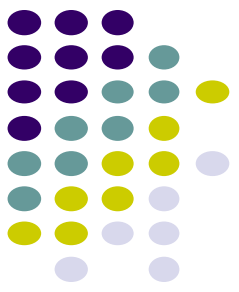
Sentence Writing

*There are at least 3 blue blocks*  
*There are 2 towers that contain yellow blocks*




Manually  
Annotated

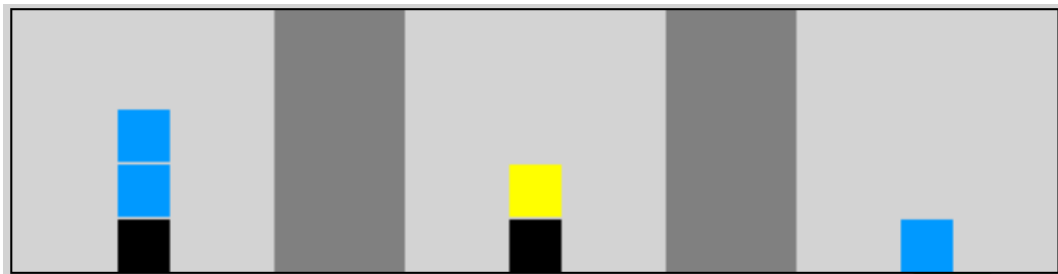
Sentence Validation

*There are at least 3 blue blocks* ✓  
*There are 2 towers that contain yellow blocks* ✗



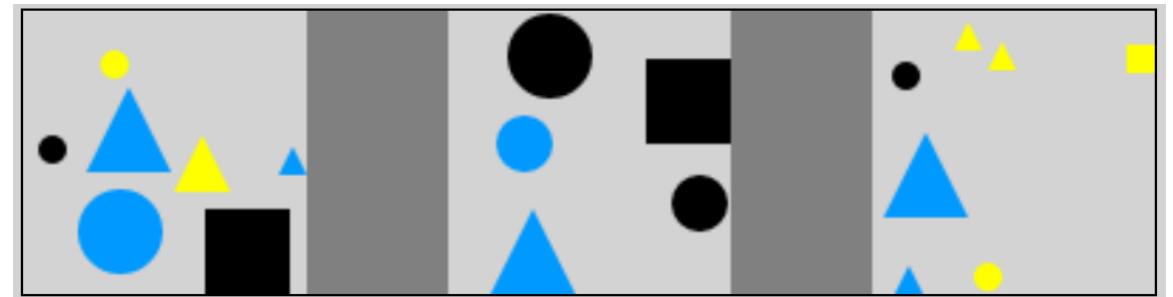
# Automatic Image Generation

- Image consists of 3 boxes, each contains 1-8 objects with the following properties:
  - **color** (black, blue, yellow) , **shape** (  ,  ,  )
  - **size** (small, medium, large) , **position** (x/y coordinate)
- Number of objects and properties are sampled uniformly
- Equal number of **tower** and **scatter** images are generated



***Tower image***

(only square objects forming towers)



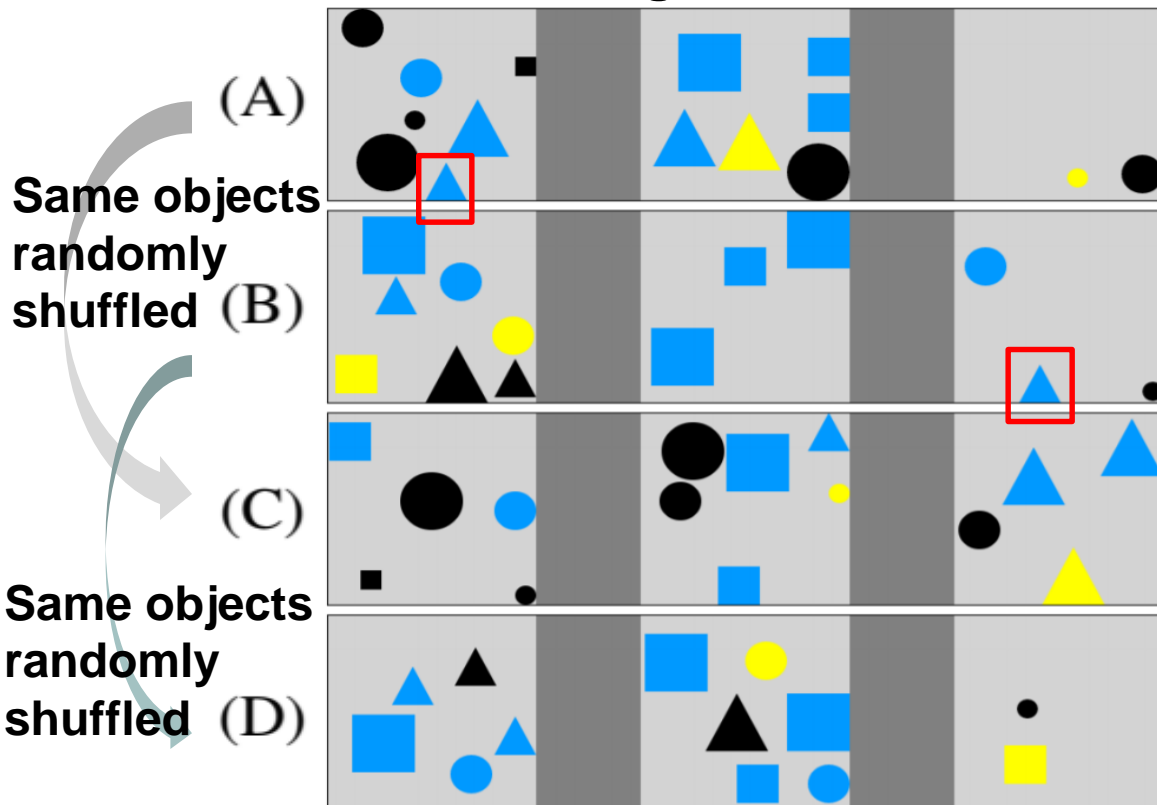
***Scatter image***

(objects are scattered around the scene)

# Sentence Writing



Annotators are presented with  
4 images at a time



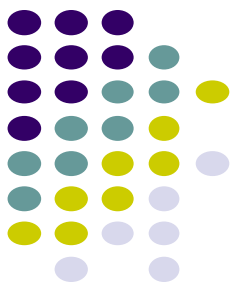
Annotation Task:

Write one sentence to meet the following requirements

- It describes (A)
- It describes (B)
- It does not describe (C)
- It does not describe (D)
- It does not mention the images explicitly (e.g., “In image A ..”)
- It does not mention the order of boxes (e.g., “In the rightmost square”)

There is no one correct sentence for this task. If you can think of more than one sentence submit only one

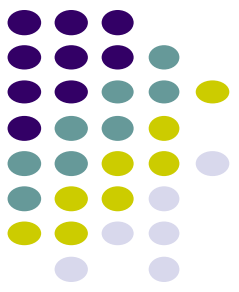
**Example: There is one blue triangle touching the bottom of one box**



# Sentence Validation

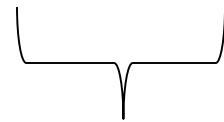
- Attach the sentence to each of the 4 images
- Randomly Permute the images and the boxes in each image

(A)		There is one blue triangle touching the bottom of one box	→	✓ or ✗
(B)		There is one blue triangle touching the bottom of one box	→	✓ or ✗
(C)		There is one blue triangle touching the bottom of one box	→	✓ or ✗
(D)		There is one blue triangle touching the bottom of one box	→	✓ or ✗

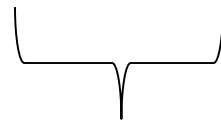


# Data Statistics

- They collected 3974 unique sentences (one sentence for 4 images)
- Dataset size =  $3974 * 4 * 6 = 95376 \approx 92244$  after pruning



sentences



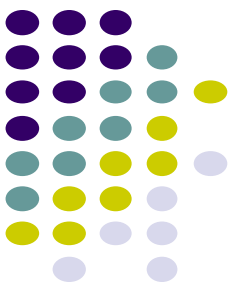
images



Box permutations

- The data is prepared by 10 annotators through crowdsourcing framework Upwork
- Total cost for annotating the data = 5,526 \$





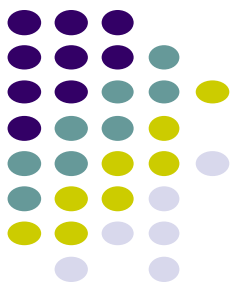
# Training Models on the data

- The paper compares several methods to perform the visual reasoning task on the proposed dataset
- The goal is to show how challenging the data is
- Three different classes of models are compared:
  - **Single modality models:** Text-only or Image-only
  - **Structured representation models:** models trained on structured representation only without image representation
    - e.g. `{"type":"square", "color":"black", "x_loc":40, "y_loc":80, "size":"20"}`, ...
  - **Image Representation models:** models trained on both image and text data (multimodal)



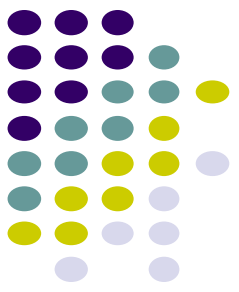
# Single Modality Models

- **Majority:** Assign the most common label (true) to all examples
- **Text-only:** Encode the sentence with RNN (LSTM + softmax)
- **Image-only:** Encode the image with CNN (3 convolutional layers + 3 feed-forward layers + softmax)



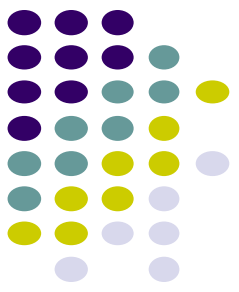
# Structured Representation Models

- **MaxEnt:** Compute Maximum entropy classifier using both:
  - Property-based features: (e.g., Topmost/lowest object in box is in this color, Whether any object is touching in any wall in any box)
  - Count-based features (e.g., the number of black triangles, number of objects touching any wall in the image)
- **MLP (Multilayer Perceptron):** use same features as MaxEnt and train a model with single-layer perceptron + softmax
- **Image Features + RNN:** use object features (color, shape, size) + RNN sentence representation as concatenated feature vector and train two layer perceptron + softmax



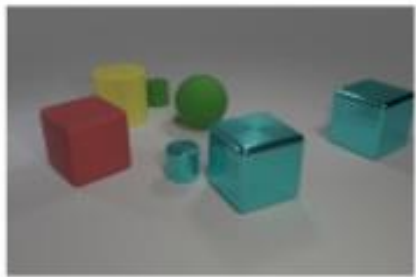
# Multimodal Models

- **CNN + RNN:** Concatenate the CNN image embedding and RNN sentence embedding, and train a multilayer perceptron with a softmax
- **NMN (Neural module networks):** neural network that is assembled dynamically by composing shallow network fragments called *modules*
  - NMNs are originally proposed for Visual Question Answering(VQA)
  - “Deep Compositional Question Answering with Neural Module Networks.” Jacob Andreas, Marcus Rohrbach, Trevor Darrell and Dan Klein. CVPR 2016.  
<https://arxiv.org/pdf/1511.02799.pdf>

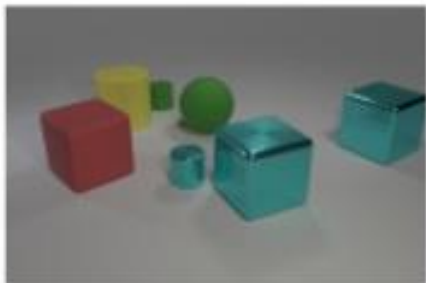


# Neural Module Networks (NMNs)

- Let's say we want to answer these two questions:
  - What color is the thing with the same size as the blue cylinder ?

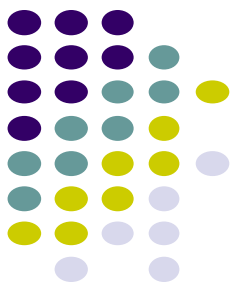


**Answer:** Green



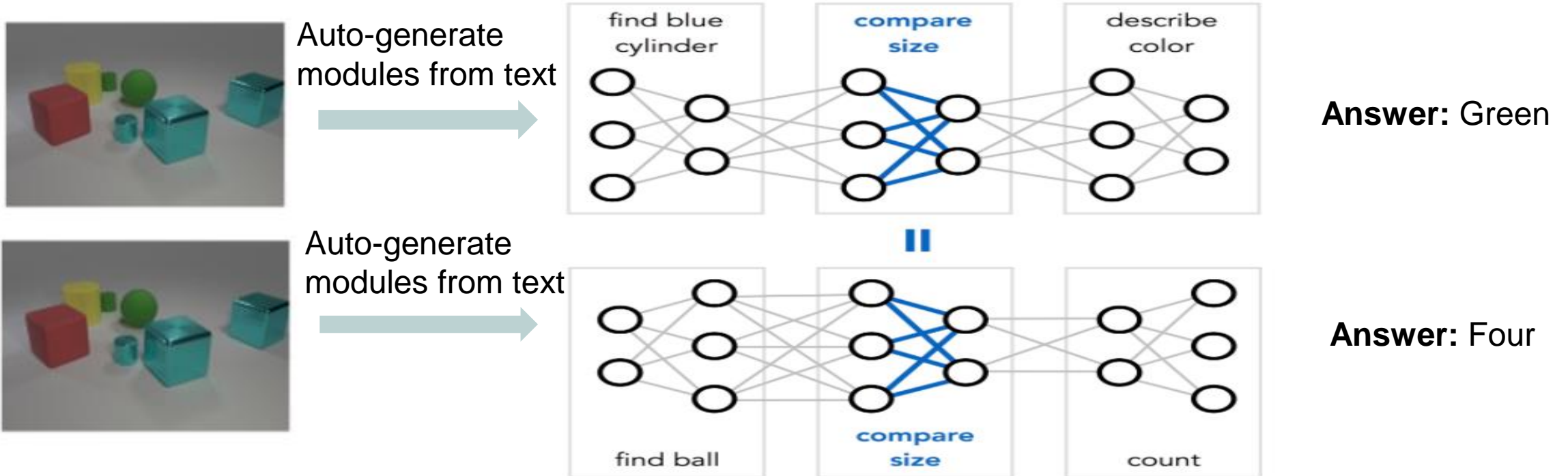
**Answer:** Four

- How many things are the same size of the ball ?

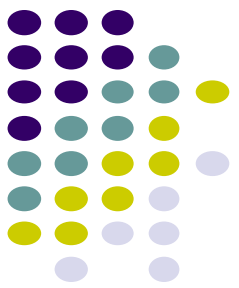


# Neural Module Networks (NMNs) Magic

- Let's say we want to answer these two questions
  - What color is the thing with the same size as the blue cylinder ?



- How many things are the same size of the ball

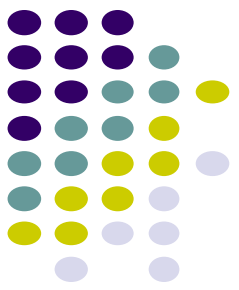


# Results

		Train	Dev	Test-P	Test-U
	Majority	56.37	55.31	56.16	55.43
	Text only	58.36 $\pm 0.6$	56.61 $\pm 0.5$	57.18 $\pm 0.6$	56.21 $\pm 0.4$
	Image Only	56.79 $\pm 1.3$	55.35 $\pm 0.1$	56.05 $\pm 0.3$	55.33 $\pm 0.3$
Structured representation	MaxEnt	99.99	68.04	67.68	67.82
	MLP	96.15 $\pm 1.3$	67.50 $\pm 0.5$	66.28 $\pm 0.4$	65.32 $\pm 0.4$
	Image features+RNN	59.71 $\pm 1.0$	57.72 $\pm 1.4$	57.62 $\pm 1.3$	56.29 $\pm 0.9$
Raw image	CNN+RNN	58.85 $\pm 0.2$	56.59 $\pm 0.3$	58.01 $\pm 0.3$	56.30 $\pm 0.6$
	NMN	98.37 $\pm 0.6$	63.06 $\pm 0.1$	66.12 $\pm 0.4$	61.99 $\pm 0.8$

- **Test-P:** publicly released test set , **Test-U:** requires submitting trained models
- NMN is the best performing model using images (accuracy is only 66.12%)
- MaxEnt is the best performing in structured representation (when disabling count-based features accuracy drops from 68% to 57%)

# Summary



- The paper introduces Cornell Natural Language Visual Reasoning (NLVR) dataset and task (<http://lic.nlp.cornell.edu/nlvr/>)
- The task requires reasoning about **colors**, **shapes**, and **quantities**
- The paper describes the process of creating the dataset (10 annotators , 5,526\$)
- The paper experiments with various and the best performance is relatively low (67%) which exemplifies the complexity of the data