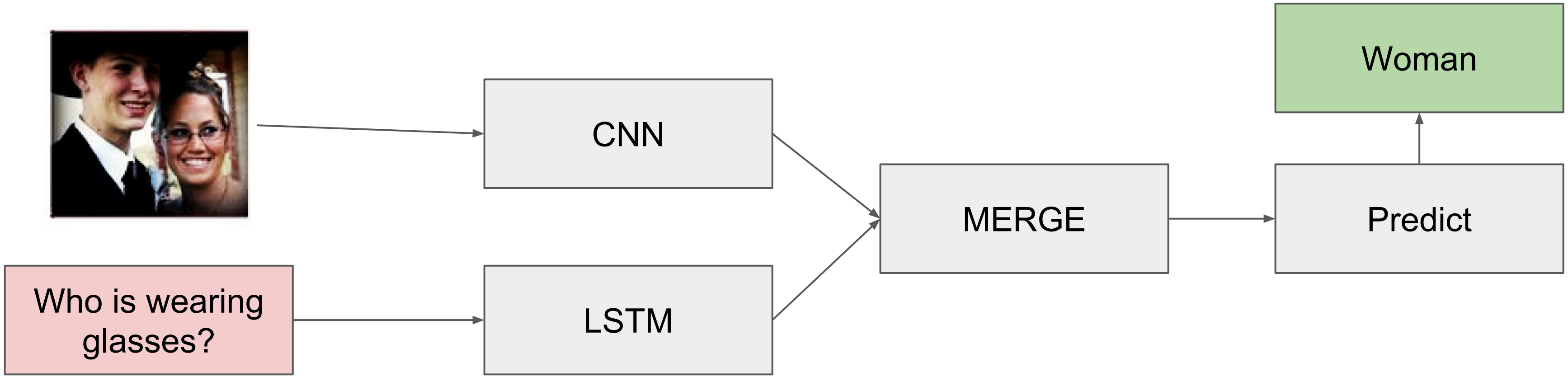
# Inferring and executing programs for Visual Reasoning

Justin Johnson, Bharath Hariharan, Laurens Maaten, Judy Hoffman,
Li Fei-Fei, C.Lawrence Zitnick, Ross Girshick
Stanford University, Facebook Research
International Conference on Computer Vision (ICCV 2017)

**CPSC 532L presentation**

Presented by:
Gursimran Singh
Borna Ghotbi
*{msimar,bgotbi}@cs.ubc.ca*

# Visual question answering

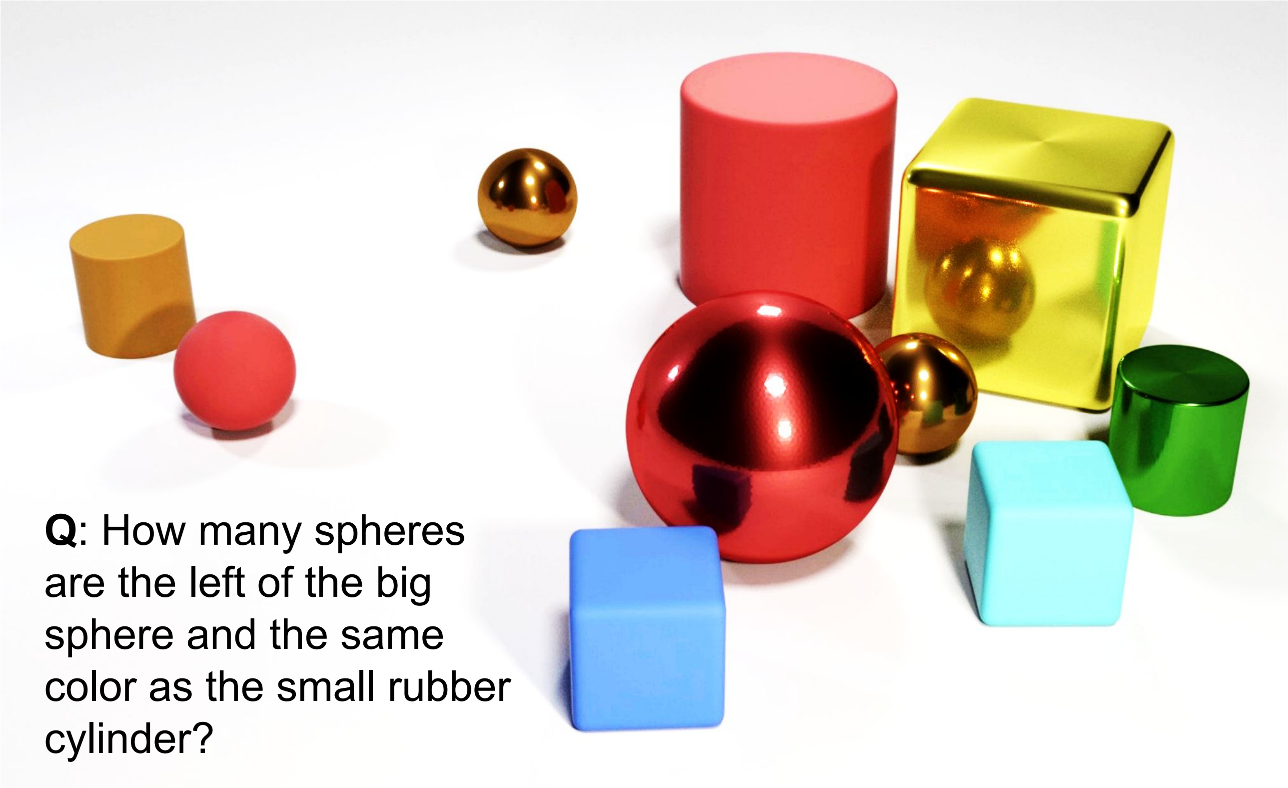



- Generalizes well to new kinds of questions
    - who is wearing spectacles; how many people?



Antol etal. Vqa: Visual question answering : Proceedings of the IEEE International Conference on Computer Vision

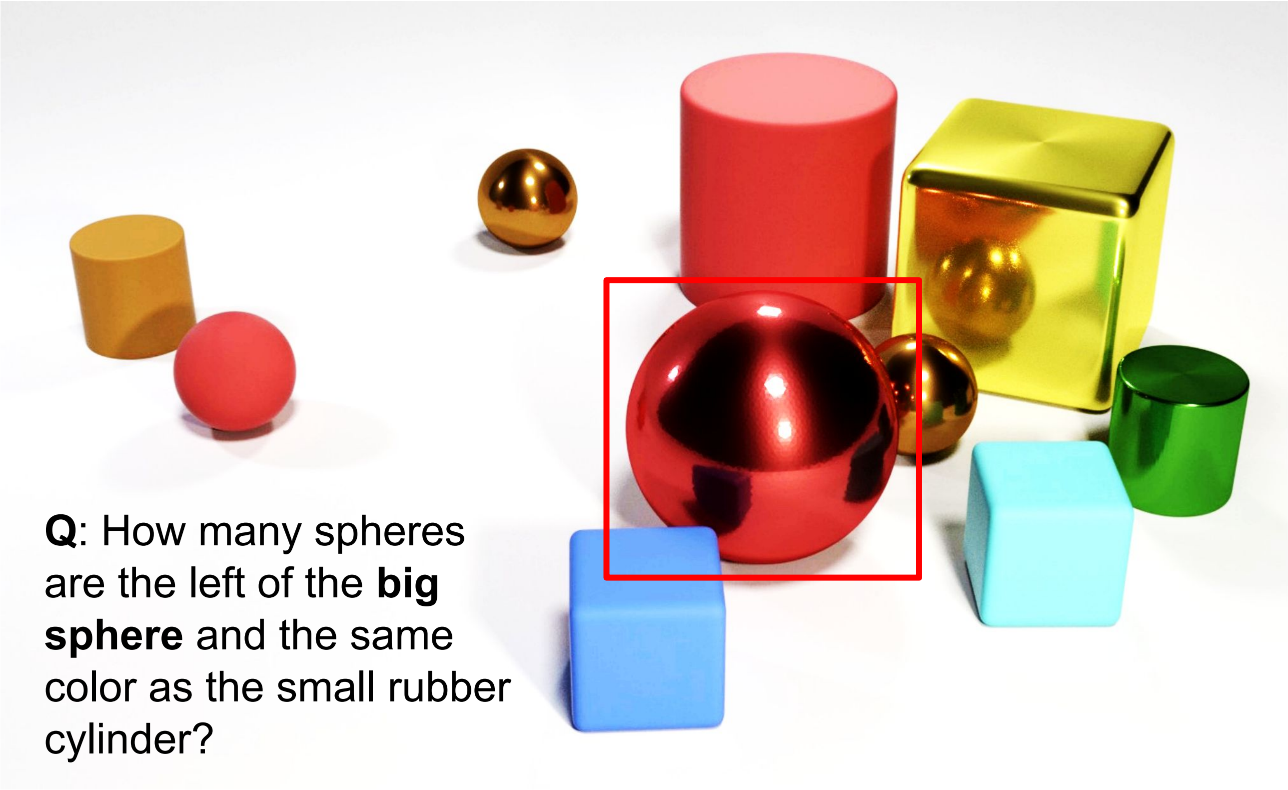
# Compositional visual reasoning



**Q**: How many spheres are the left of the big sphere and the same color as the small rubber cylinder?

Johnson, Justin, et al. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. *(CVPR), 2017*

# Compositional visual reasoning

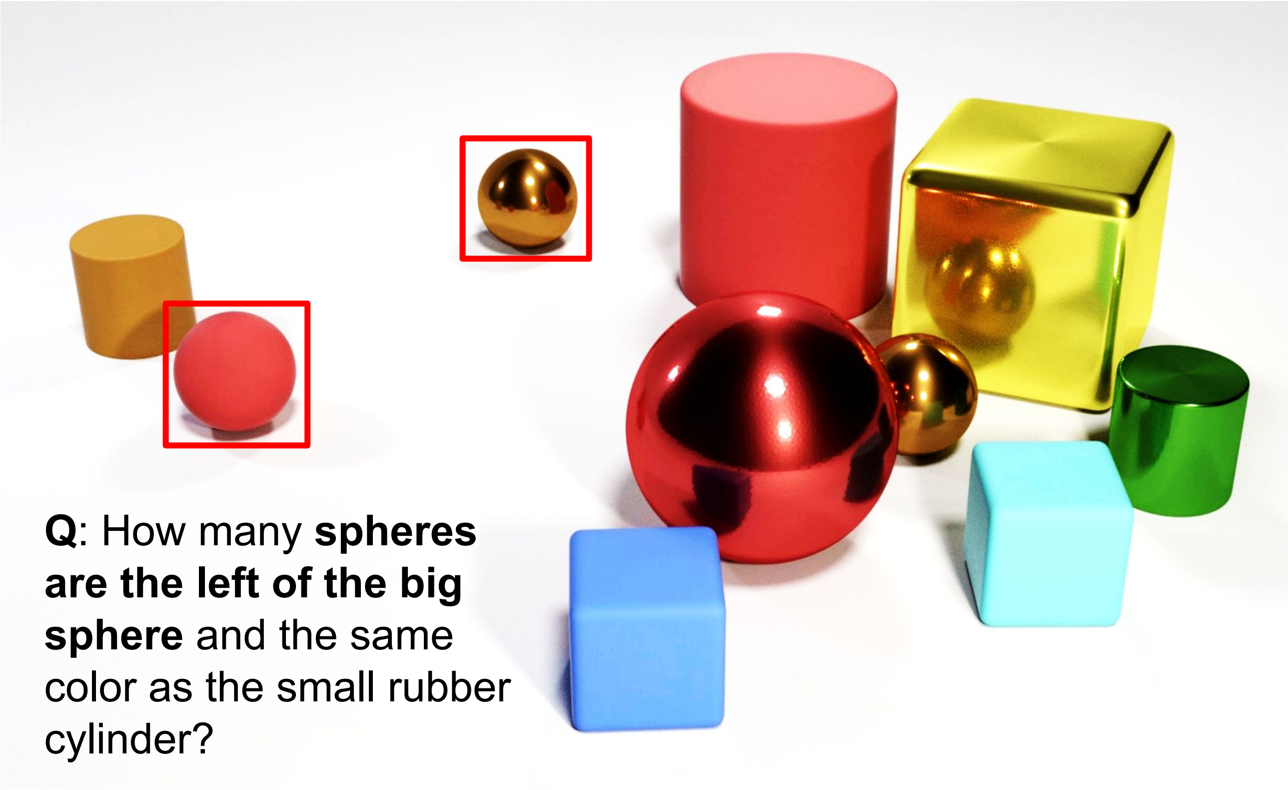


Identify big sphere

**Q**: How many spheres are the left of the **big sphere** and the same color as the small rubber cylinder?

Johnson, Justin, et al. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. *(CVPR), 2017*

# Compositional visual reasoning



**Q**: How many **spheres are the left of the big sphere** and the same color as the small rubber cylinder?

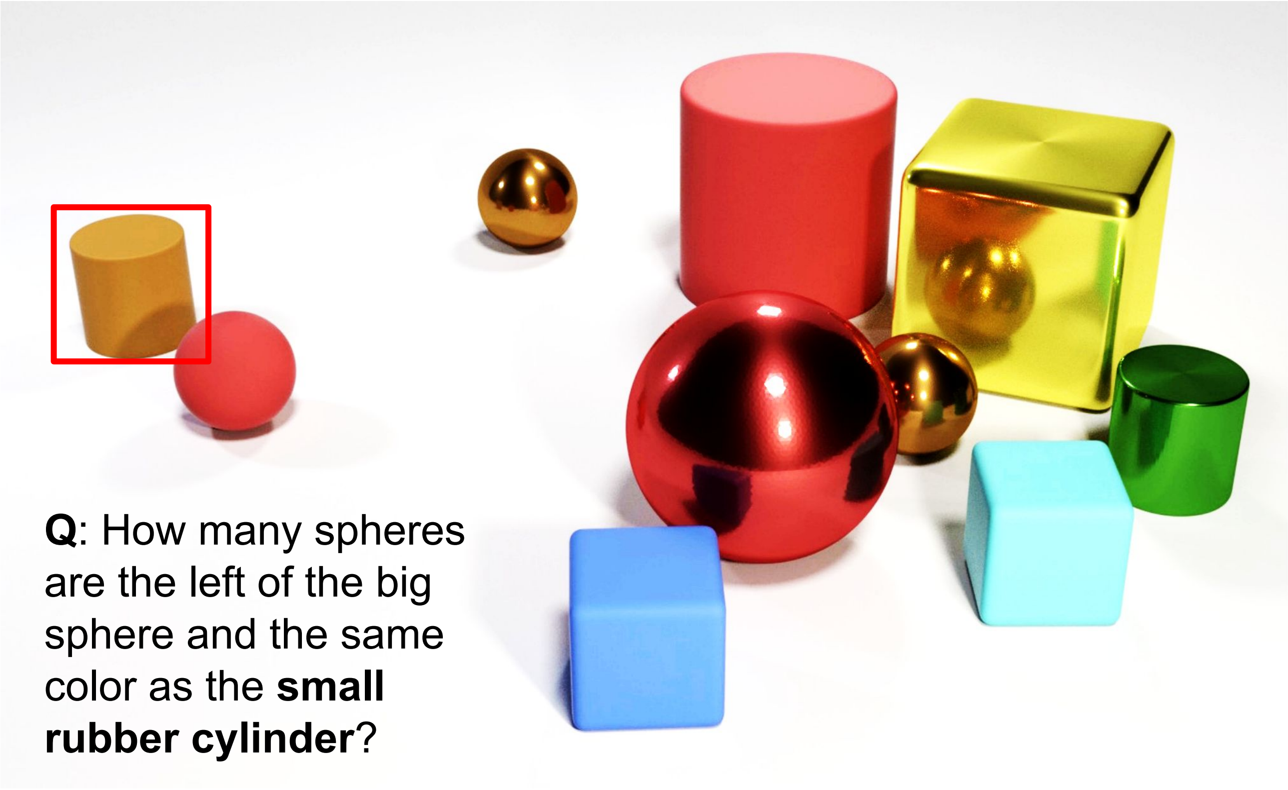
Identify big sphere

↓

Spheres on left

Johnson, Justin, et al. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. *(CVPR), 2017*

# Compositional visual reasoning



**Q**: How many spheres are the left of the big sphere and the same color as the **small rubber cylinder**?

Identify big sphere

↓

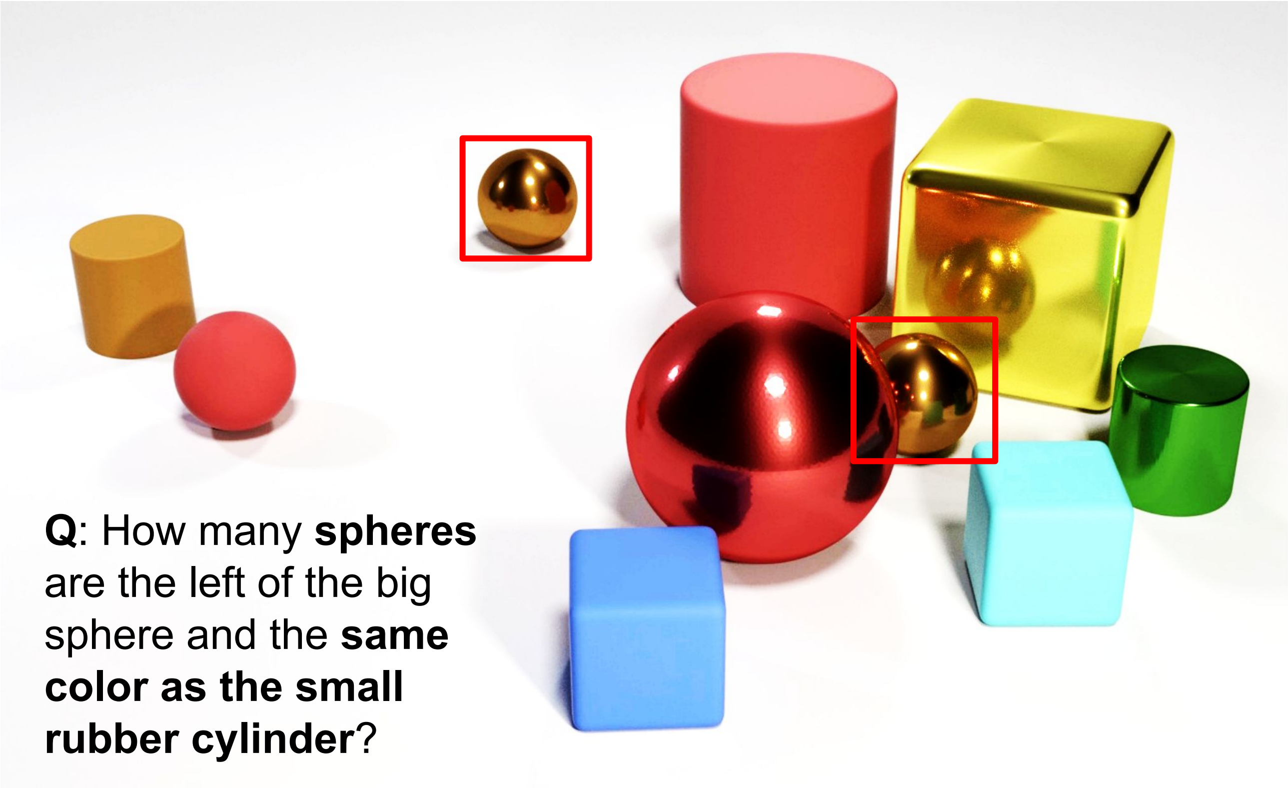
Spheres on left

↓

Rubber cylinder

Johnson, Justin, et al. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. *(CVPR), 2017*

# Compositional visual reasoning



**Q**: How many **spheres** are the left of the big sphere and the **same color as the small rubber cylinder**?

Identify big sphere

↓

Spheres on left

↓

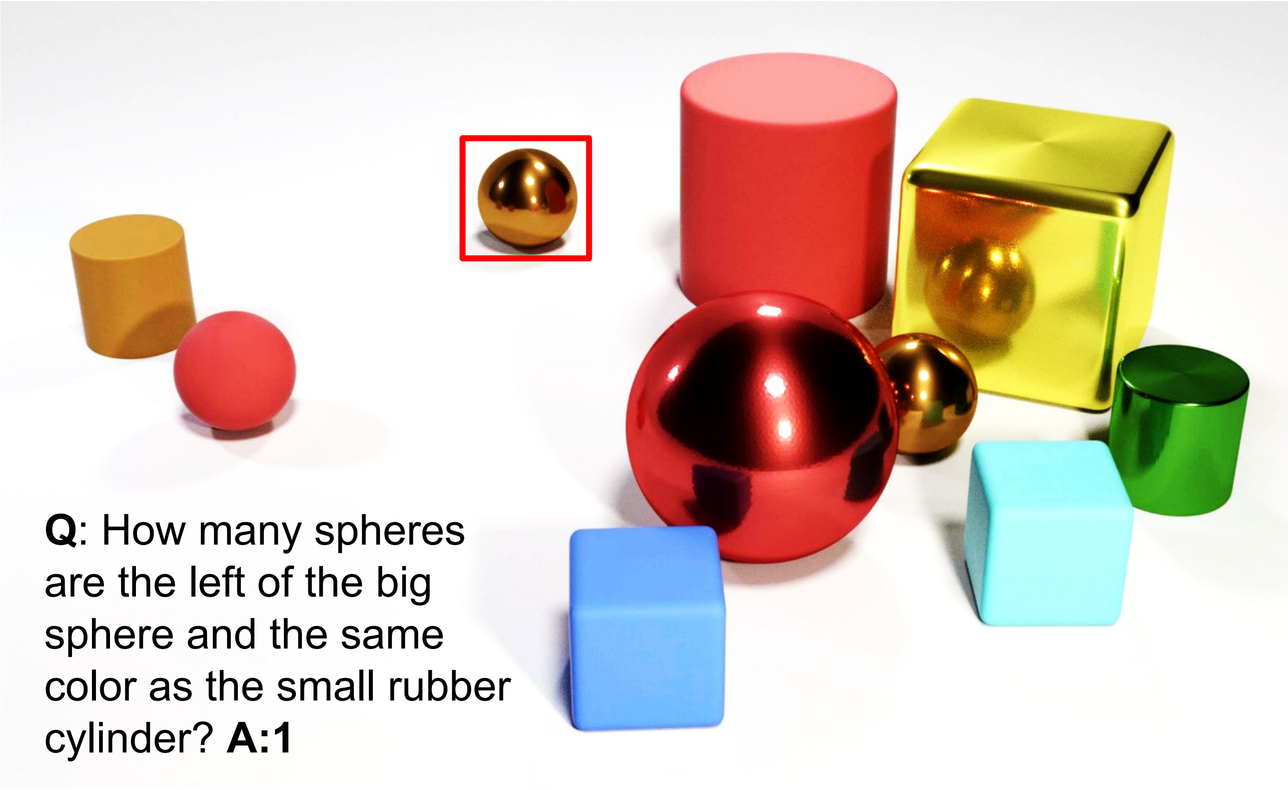
Rubber cylinder

↓

Sphere of same color

Johnson, Justin, et al. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. *(CVPR), 2017*

# Compositional visual reasoning



**Q**: How many spheres are the left of the big sphere and the same color as the small rubber cylinder? **A:1**

Identify big sphere

↓

Spheres on left

↓

Rubber cylinder

↓

Sphere of same color

↓

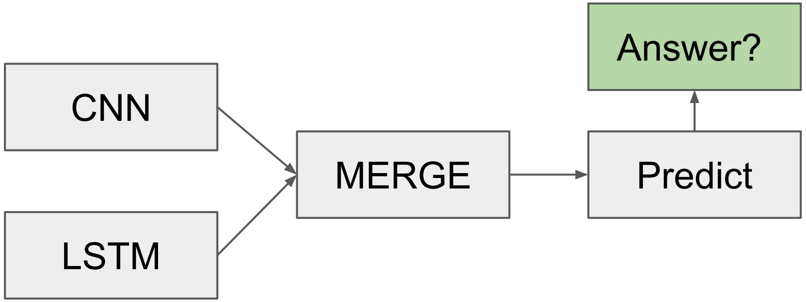
Count



Johnson, Justin, et al. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. *(CVPR), 2017*

# Standard VQA?

**Q**: How many spheres are the left of the big sphere and the same color as the small rubber cylinder?



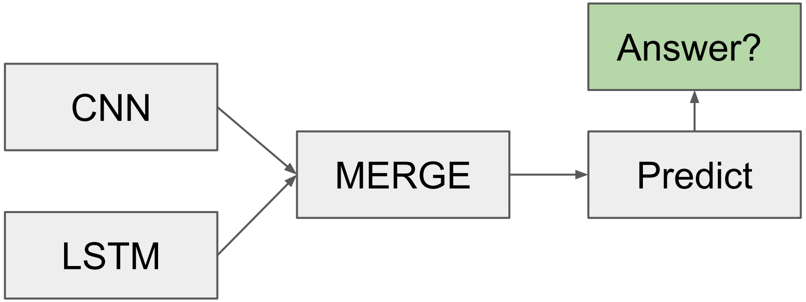

**LIMITATIONS**

- Can’t model complex questions

# Standard VQA?

**Q**: How many spheres are the **left** of the big sphere and the same color as the small rubber cylinder?

**Q**: How many spheres are the **right** of the big sphere and the same color as the small rubber cylinder?



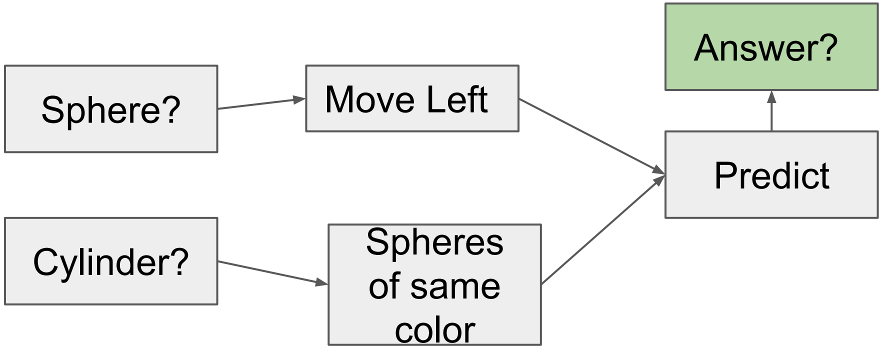

**LIMITATIONS**

- Can’t model complex questions
- Lacks composition

# Standard VQA?

**Q**: How many spheres are the **left** of the big sphere and the same color as the small rubber cylinder?

**Q**: How many spheres are the **right** of the big sphere and the same color as the small rubber cylinder?




**LIMITATIONS**

- Can't model complex questions
- Lacks composition

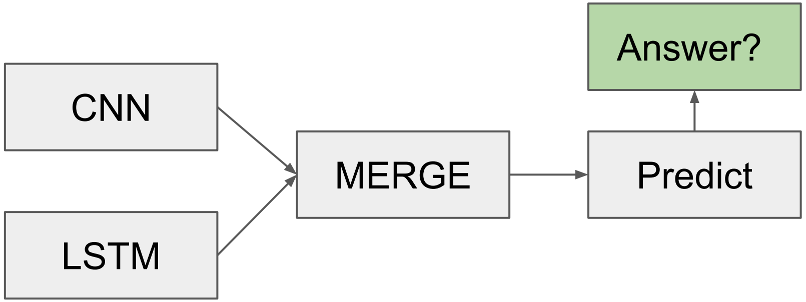
**Decompose the network into multiple modules**

# Standard VQA?

**Q**: How many spheres are the **left** of the big sphere and the same color as the small rubber cylinder?

**Q**: How many spheres are the **right** of the big sphere and the same color as the small rubber cylinder?

**Q**: How many objects are either red cylinders or metal objects?




**LIMITATIONS**

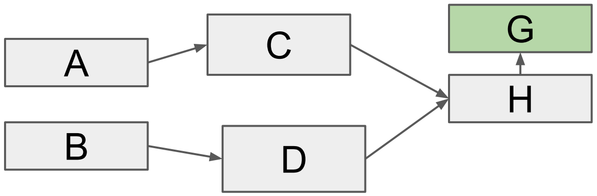
- Can’t model complex questions
- Lacks composition
- Uses same structure

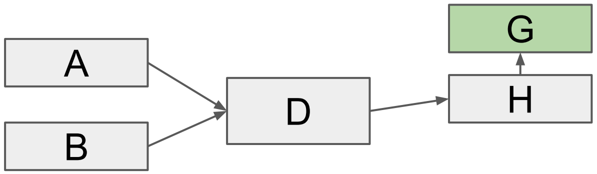
# Standard VQA?

**Q**: How many spheres are the **left** of the big sphere and the same color as the small rubber cylinder?

**Q**: How many spheres are the **right** of the big sphere and the same color as the small rubber cylinder?

**Q**: How many objects are either red cylinders or metal objects?




**Use separate networks for each question**

**LIMITATIONS**

- Can't model complex questions
- Lacks composition
- Uses same structure

**Solution**

- Use composition and structure

# Instead: consider a compositional model

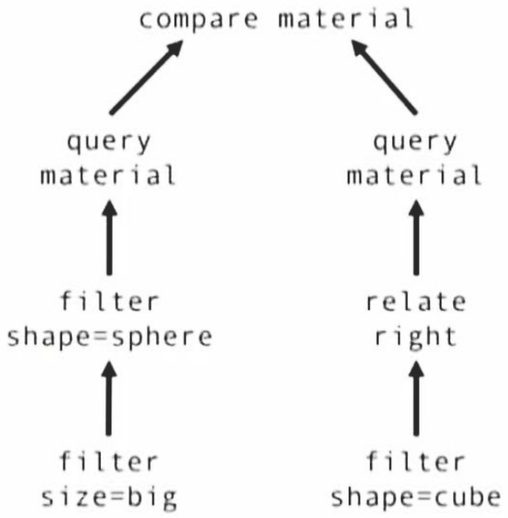
**Q**: How many spheres are the left of the big sphere and the same color as the small rubber cylinder?

**Q**: How many spheres are the right of the big sphere and the same color as the small rubber cylinder?

**Q**: Is the big sphere the same material as the thing on the right of the cube?

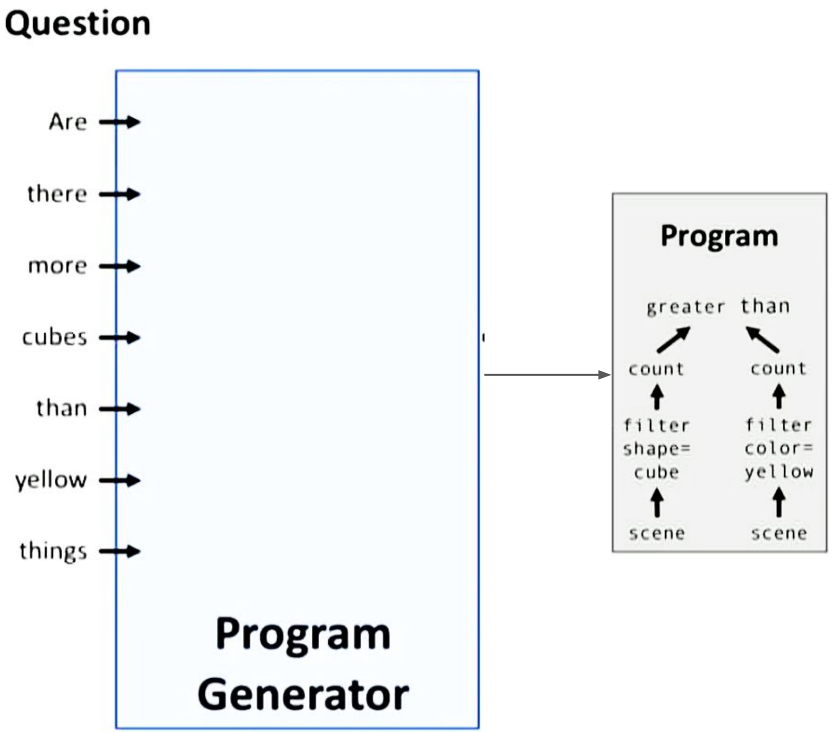
**Common operations**

Attributes identification
Counting objects
Comparisons
Spatial relationships
Logical operations



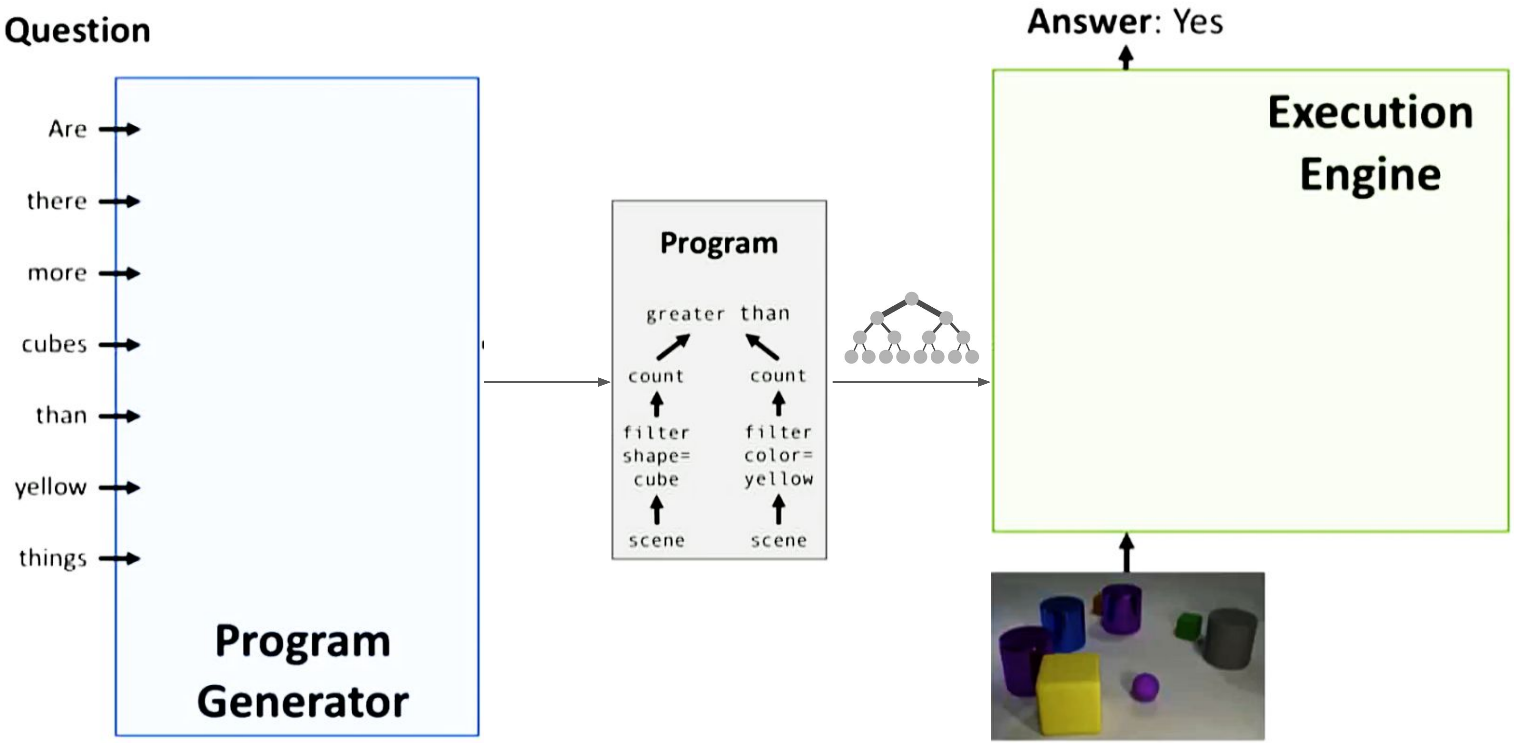

**Network architecture corresponding to the third question**

# Overview of approach

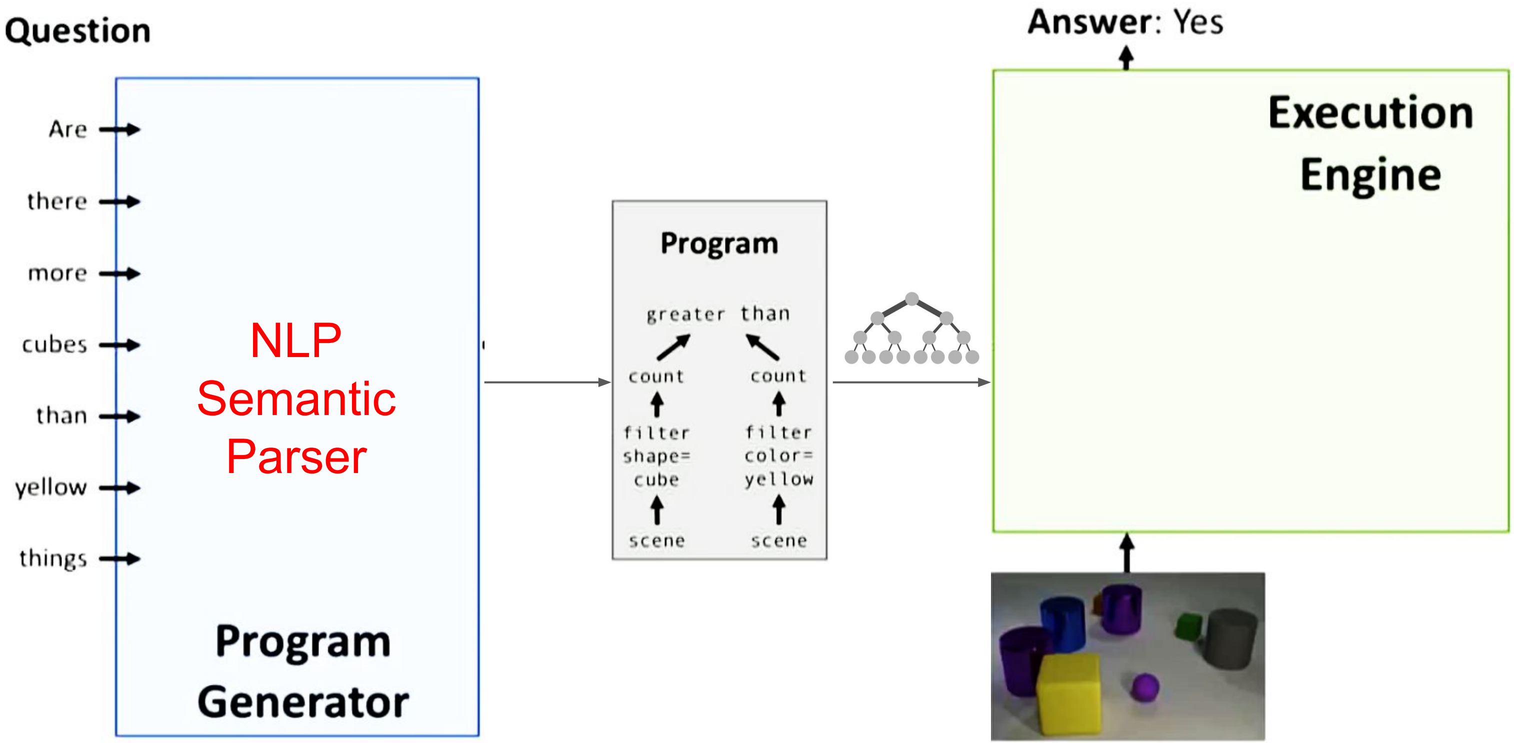



Graphics take from -> https://www.youtube.com/watch?v=3pCLma2FqSk

# Overview of approach

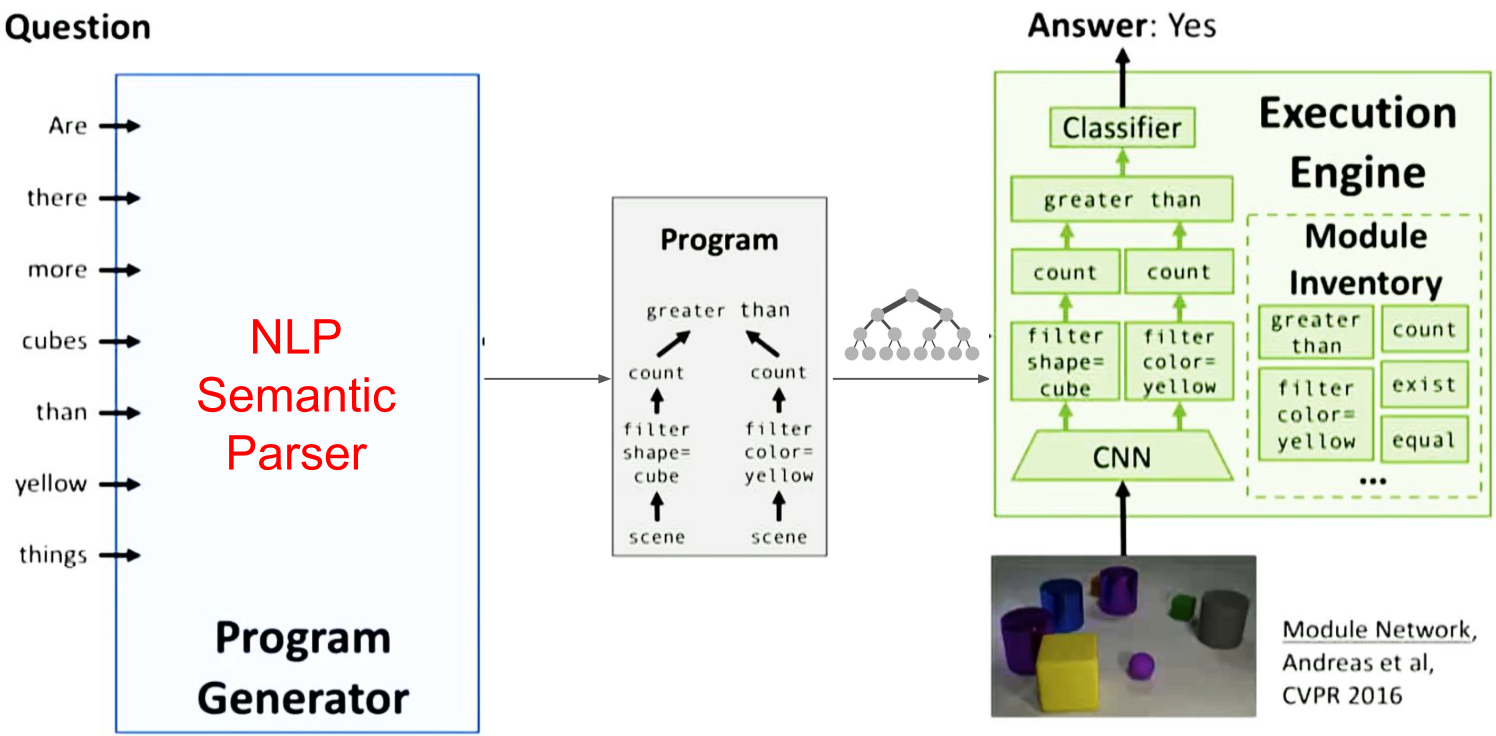


# Module networks




Graphics take from -> https://www.youtube.com/watch?v=3pCLma2FqSk

# Module networks

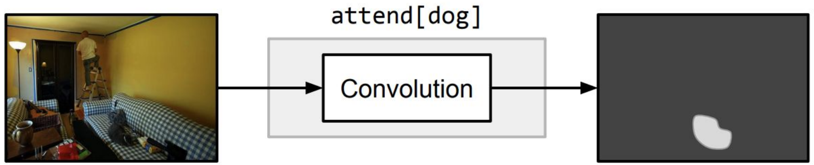



Graphics take from -> https://www.youtube.com/watch?v=3pCLma2FqSk

# Modules recap

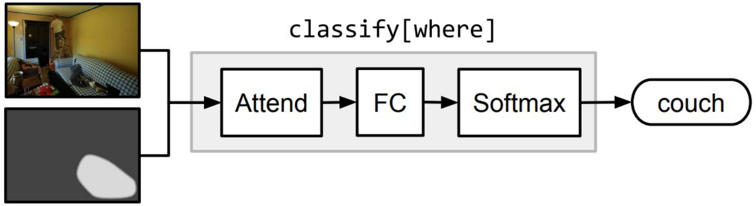
**Attention**

$\texttt{attend} : \mathit{Image} \to \mathit{Attention}$



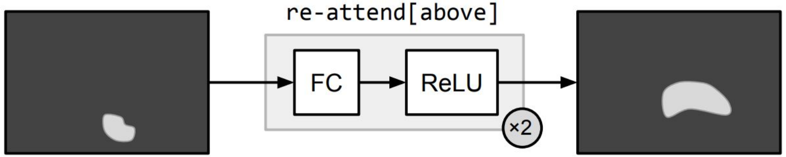

**Classification**

$\texttt{classify} : \mathit{Image} \times \mathit{Attention} \to \mathit{Label}$



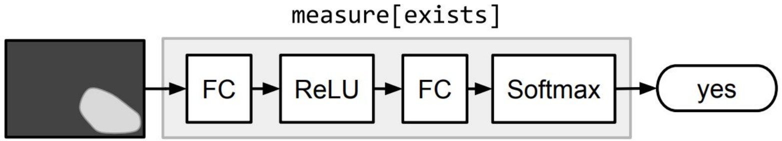

**Re-attention**

$\texttt{re-attend} : \mathit{Attention} \to \mathit{Attention}$



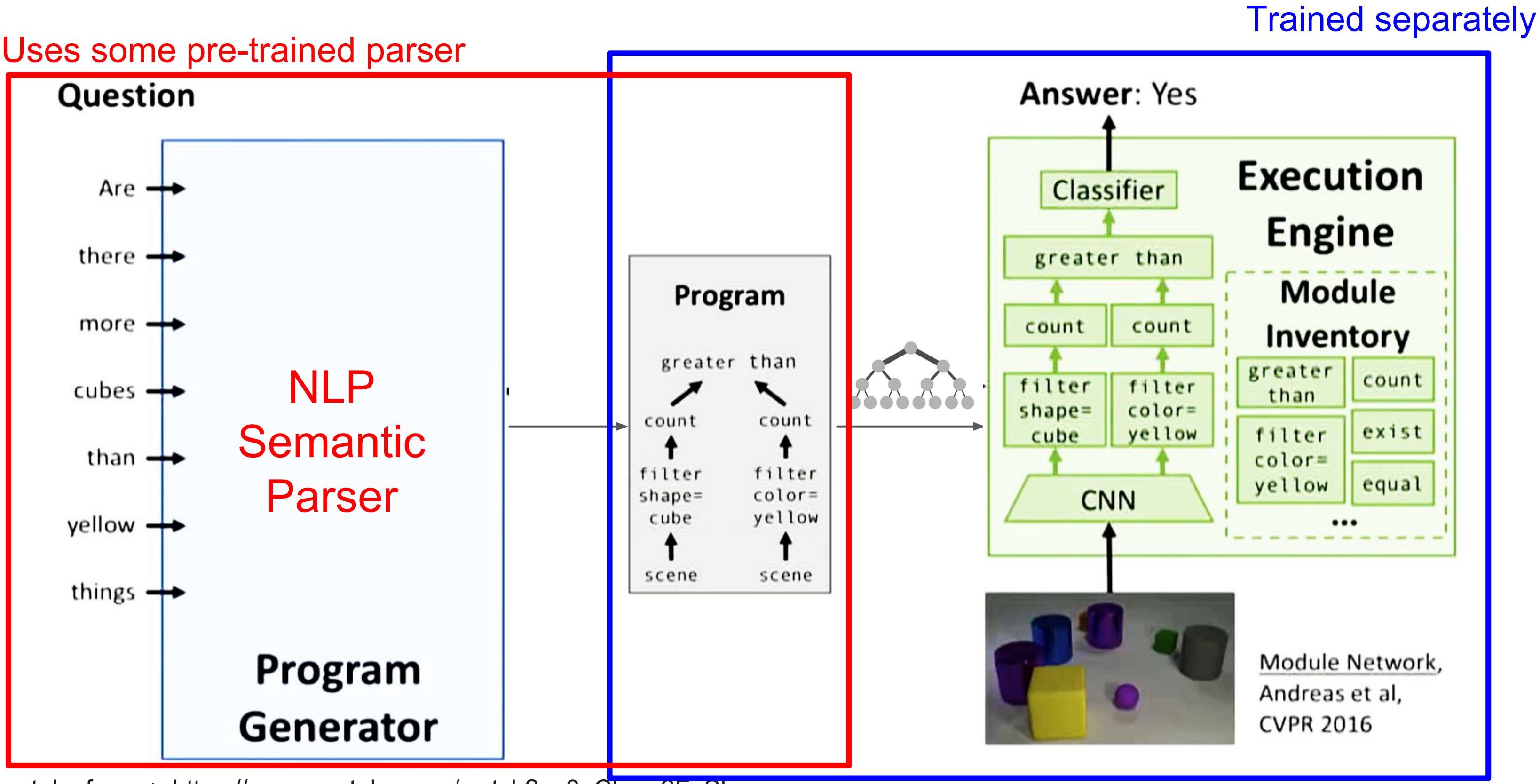

**Measurement**

$\texttt{measure} : \mathit{Attention} \to \mathit{Label}$




Andreas etal; Deep Compositional Question Answering with Neural Module Networks: arxiv 2017

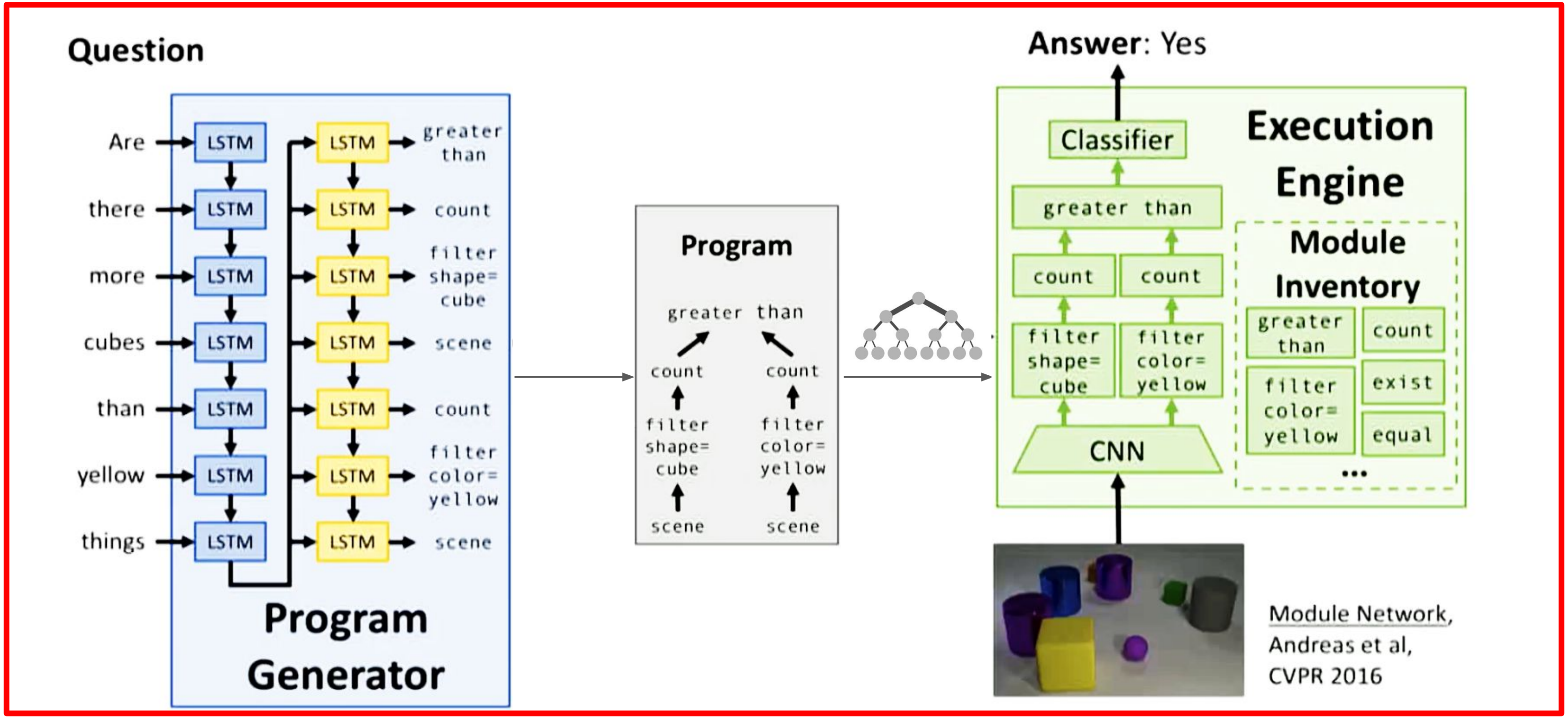
# Module networks - limitations




Graphics take from -> https://www.youtube.com/watch?v=3pCLma2FqSk

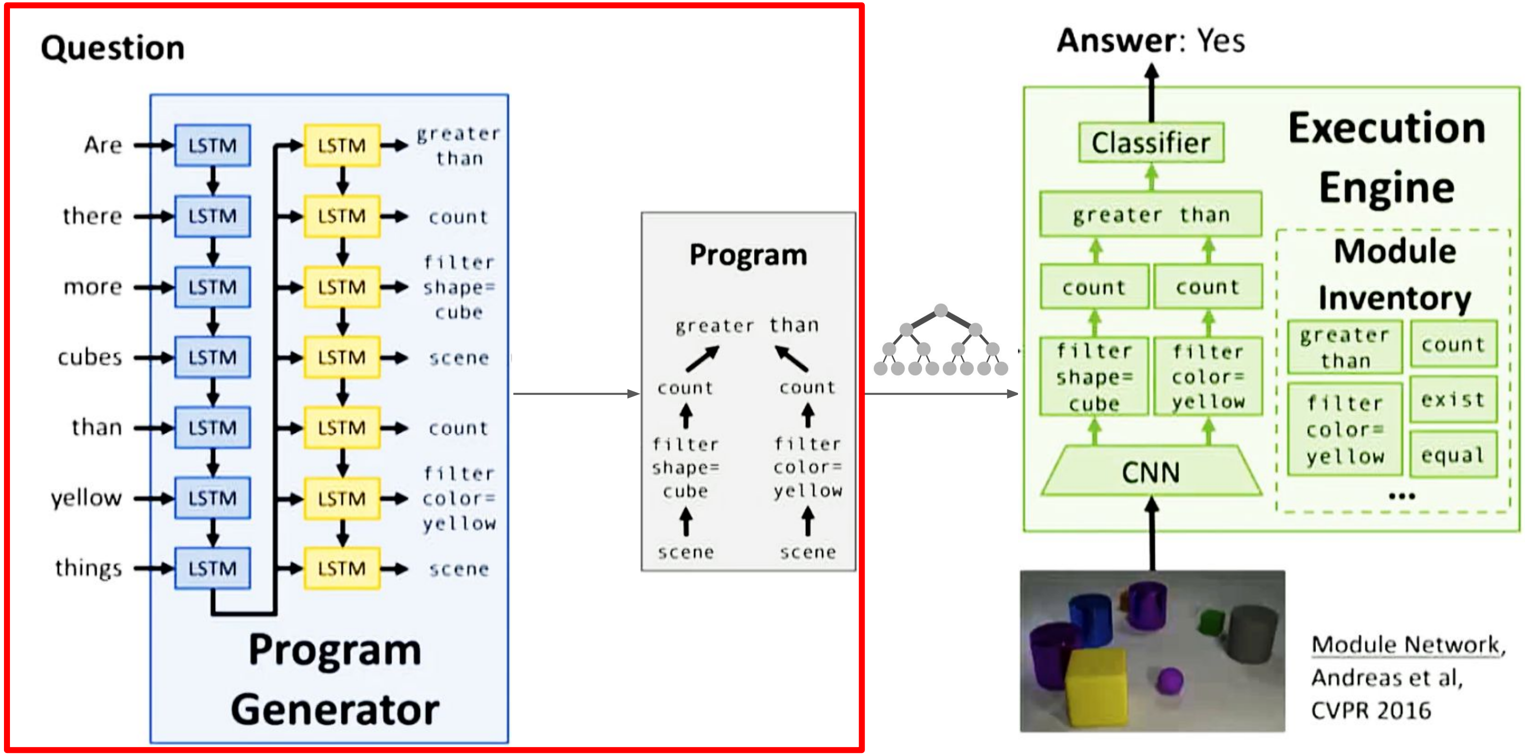
# Inferring and executing programs

Trained end-end!!!




Graphics take from -> https://www.youtube.com/watch?v=3pCLma2FqSk

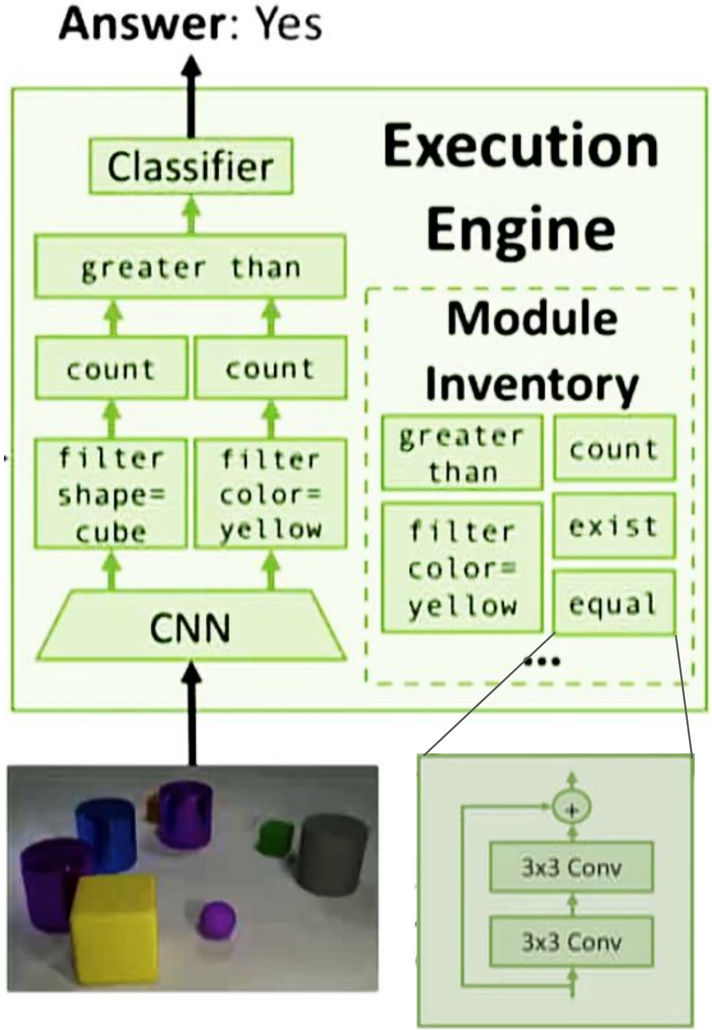
# Inferring and executing programs




Module Network, Andreas et al, CVPR 2016

Graphics take from -> https://www.youtube.com/watch?v=3pCLma2FqSk

# Execution engine



# Modules architectures

| Layer | Output size |
|---|---|
| Input image | $3 \times 224 \times 224$ |
| ResNet-101 [14] conv4_6 | $1024 \times 14 \times 14$ |
| Conv($3 \times 3$, $1024 \rightarrow 128$) | $128 \times 14 \times 14$ |
| ReLU | $128 \times 14 \times 14$ |
| Conv($3 \times 3$, $128 \rightarrow 128$) | $128 \times 14 \times 14$ |
| ReLU | $128 \times 14 \times 14$ |

a) Visual feature extraction

| Index | Layer | Output size |
|---|---|---|
| (1) | Previous module output | $128 \times 14 \times 14$ |
| (2) | Previous module output | $128 \times 14 \times 14$ |
| (3) | Concatenate (1) and (2) | $256 \times 14 \times 14$ |
| (4) | Conv($1 \times 1$, $256 \rightarrow 128$) | $128 \times 14 \times 14$ |
| (5) | ReLU | $128 \times 14 \times 14$ |
| (6) | Conv($3 \times 3$, $128 \rightarrow 128$) | $128 \times 14 \times 14$ |
| (7) | ReLU | $128 \times 14 \times 14$ |
| (8) | Conv($3 \times 3$, $128 \rightarrow 128$) | $128 \times 14 \times 14$ |
| (9) | Residual: Add (5) and (8) | $128 \times 14 \times 14$ |
| (10) | ReLU | $128 \times 14 \times 14$ |

b.2) Binary modules

| Index | Layer | Output size |
|---|---|---|
| (1) | Previous module output | $128 \times 14 \times 14$ |
| (2) | Conv($3 \times 3$, $128 \rightarrow 128$) | $128 \times 14 \times 14$ |
| (3) | ReLU | $128 \times 14 \times 14$ |
| (4) | Conv($3 \times 3$, $128 \rightarrow 128$) | $128 \times 14 \times 14$ |
| (5) | Residual: Add (1) and (4) | $128 \times 14 \times 14$ |
| (6) | ReLU | $128 \times 14 \times 14$ |

b.1) Unary modules

| Layer | Output size |
|---|---|
| Final module output | $128 \times 14 \times 14$ |
| Conv($1 \times 1$, $128 \rightarrow 512$) | $512 \times 14 \times 14$ |
| ReLU | $512 \times 14 \times 14$ |
| MaxPool($2 \times 2$, stride 2) | $512 \times 7 \times 7$ |
| FullyConnected($512 \cdot 7 \cdot 7 \rightarrow 1024$) | 1024 |
| ReLU | 1024 |
| FullyConnected($1024 \rightarrow \lvert\mathcal{A}\rvert$) | $\lvert\mathcal{A}\rvert$ |

d) Classifier

# What do the modules learn?

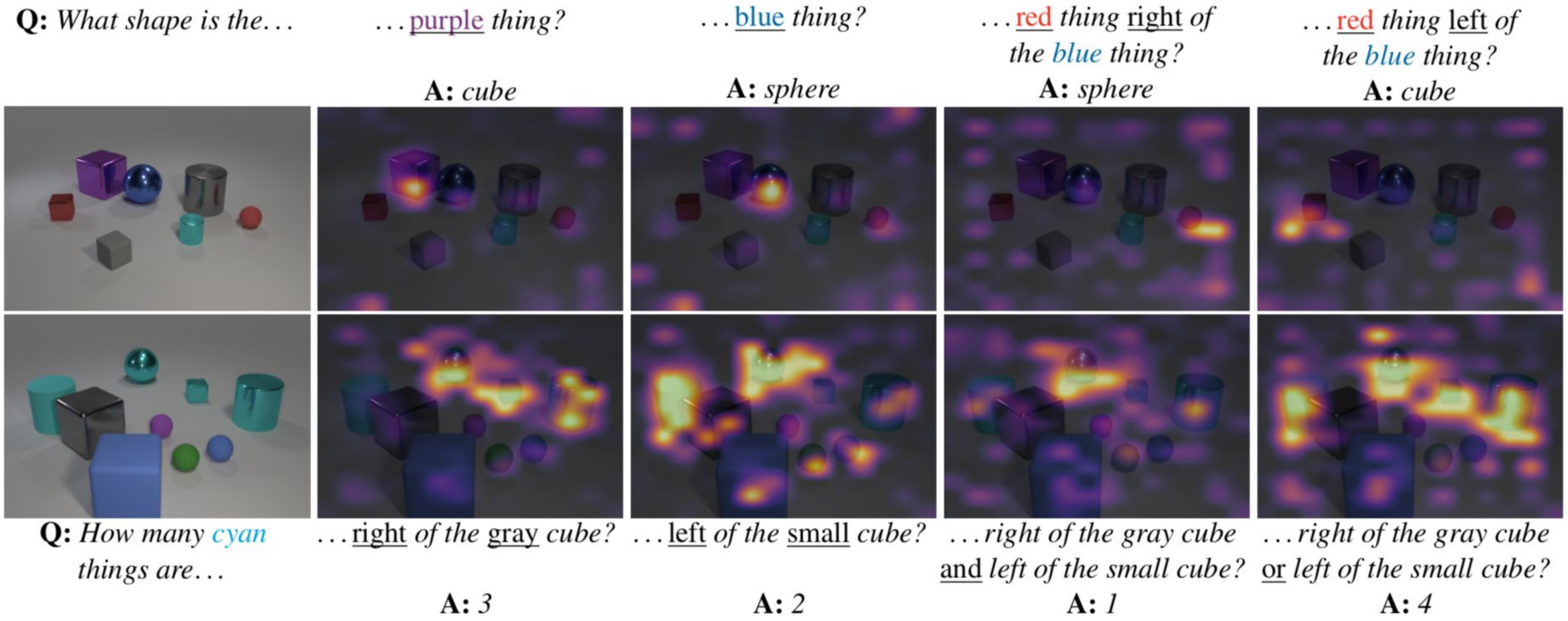
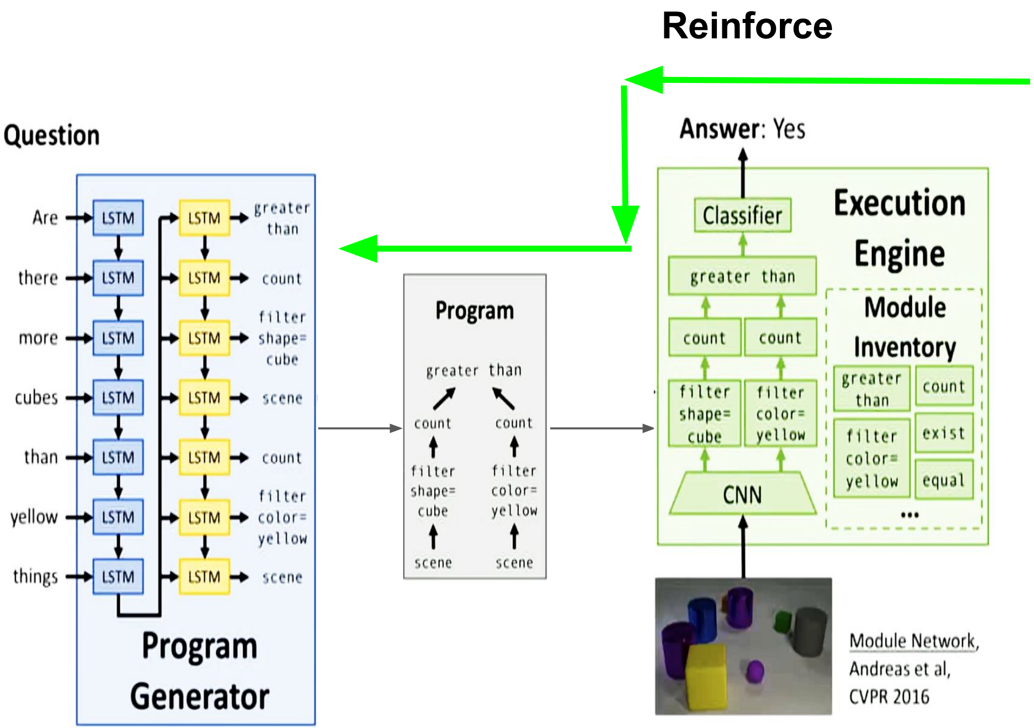



Figure 3. Visualizations of the norm of the gradient of the sum of the predicted answer scores with respect to the final feature map. From left to right, each question adds a module to the program; the new module is *underlined* in the question. The visualizations illustrate which objects the model attends to when performing the reasoning steps for question answering. Images are from the validation set.

# Training

- Train Program Generator
- Freeze Program Generator, Train Execution Engine
- Finetune




Module Network, Andreas et al, CVPR 2016

# Clever dataset

A training set of 70,000 images and 699,989 questions

● A validation set of 15,000 images and 149,991 questions

● A test set of 15,000 images and 14,988 questions

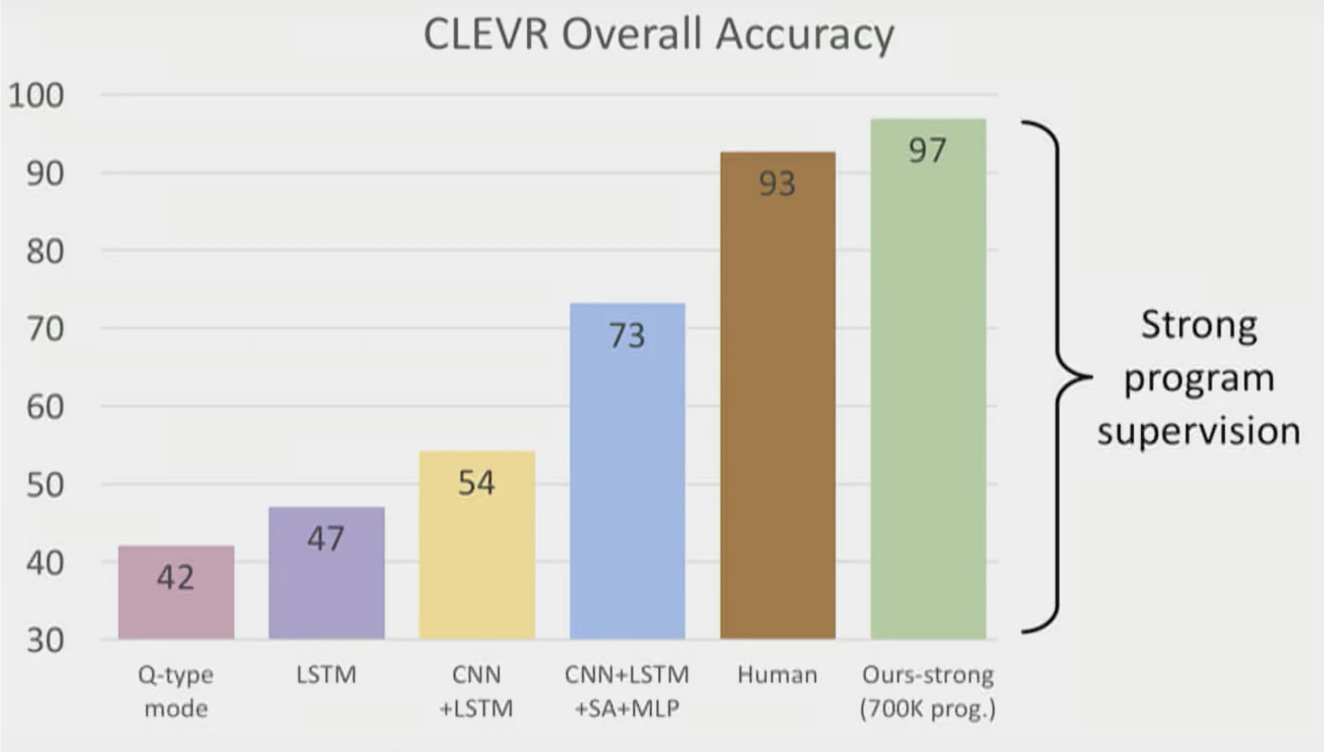
● Answers for all train and val questions

● Scene graph annotations for train and val images giving ground-truth

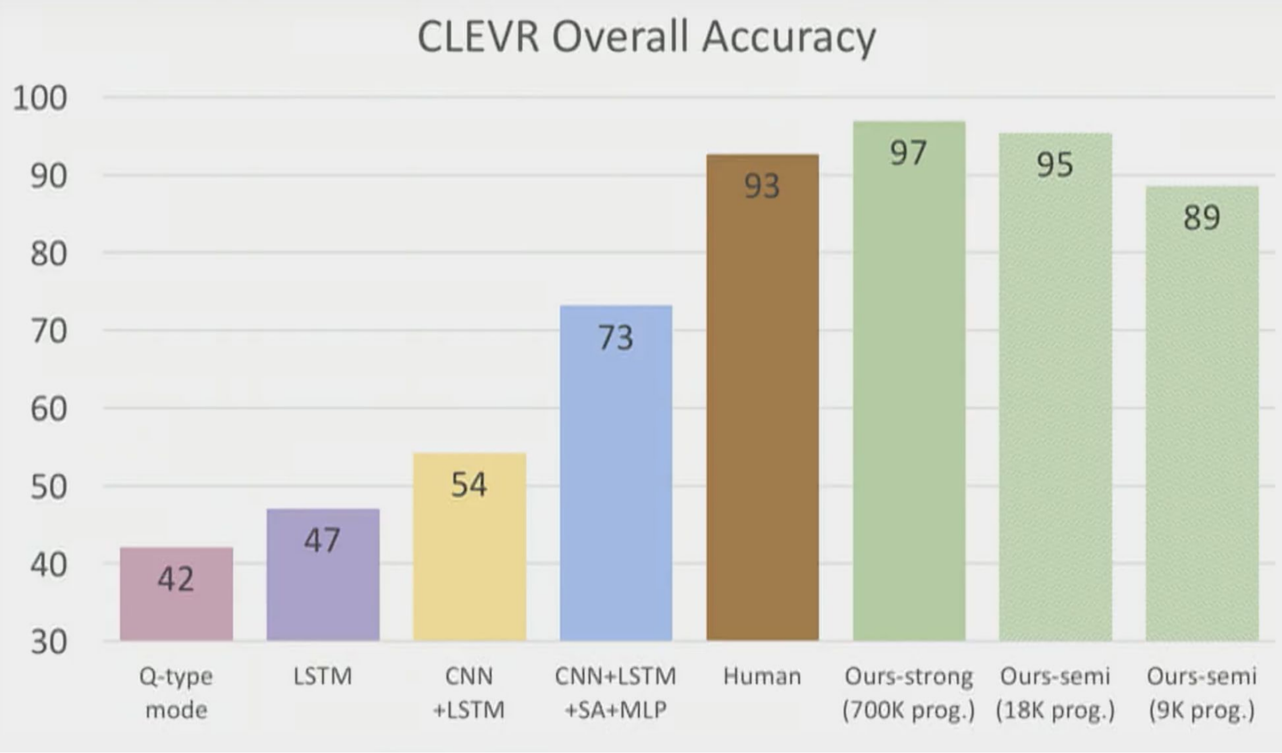
locations, attributes, and relationships for objects

● Objects can be cubes, cylinders and spheres.

# Experiments: Baselines



# Experiments: Strongly and semi-supervised learning



# Experiments

## Generalizing to new attribute combinations

### Compositional Generalization Test (CoGenT)

This data was used in Section 4.7 of the paper to study the ability of models to recognize novel combinations of attributes at test-time. The data is generated in two different conditions:

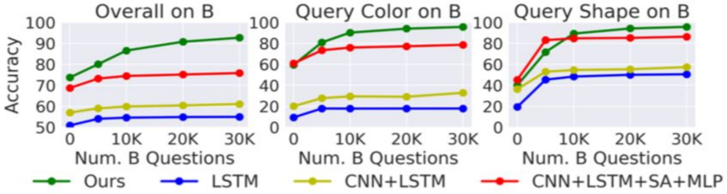
**Condition A**

- Cubes are gray, blue, brown, or yellow
- Cylinders are red, green, purple, or cyan
- Spheres can have any color

**Condition B**

- Cubes are red, green, purple, or cyan
- Cylinders are gray, blue, brown, or yellow
- Spheres can have any color

| | Train A | | Finetune B | |
|---|---|---|---|---|
| Method | A | B | A | B |
| LSTM | 55.2 | 50.9 | 51.5 | 54.9 |
| CNN+LSTM | 63.7 | 57.0 | 58.3 | 61.1 |
| CNN+LSTM+SA+MLP | 80.3 | 68.7 | 75.7 | 75.8 |
| Ours (18K prog.) | **96.6** | **73.7** | **76.1** | **92.7** |



# Experiments

## Generalizing to new question types

**Short**: all questions which their questions family has a mean program length less than 16

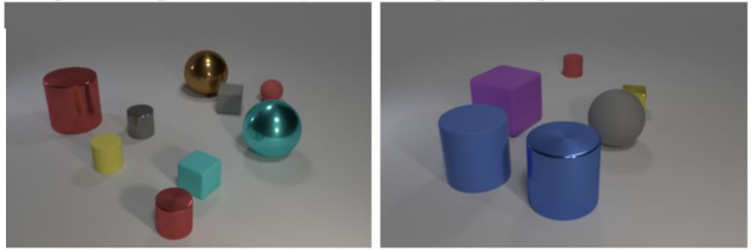
**Long**: otherwise

| Method | Train Short: Short | Train Short: Long | Finetune Both: Short | Finetune Both: Long |
|---|---|---|---|---|
| LSTM | 46.4 | 48.6 | 46.5 | 49.9 |
| CNN+LSTM | 54.0 | 52.8 | 54.3 | 54.2 |
| CNN+LSTM+SA+MLP | 74.2 | **64.3** | 74.2 | 67.8 |
| Ours (25K prog.) | **95.9** | 55.3 | **95.6** | **77.8** |

Table 2. Question answering accuracy on short and long CLEVR questions. **Left columns**: Models trained only on short questions; our model uses 25K ground-truth short programs. **Right columns**: Models trained on both short and long questions. Our model is trained on short questions then finetuned on the entire dataset; no ground-truth programs are used during finetuning.

**Ground-truth question**:
*Is the number of matte blocks in front of the small yellow cylinder greater than the number of red rubber spheres to the left of the large red shiny cylinder?*
**Program length:** 20 **A:** *yes* ✓

**Ground-truth question**:
*How many objects are big rubber objects that are in front of the big gray thing or large rubber things that are in front of the large rubber sphere?*
**Program length:** 16 **A:** *1* ✓



**Predicted program** (translated):
*Is the number of matte blocks in front of the small yellow cylinder greater than the number of large red shiny cylinders?*
**Program length:** 15 **A:** *no* ✗

**Predicted program** (translated):
*How many objects are big rubber objects in front of the big gray thing or large rubber spheres?*
**Program length:** 12 **A:** *2* ✗

# Experiments

## The CLEVR-Humans Dataset

- Use of questions that are hard to answer for a “smart robot”
- Filtered questions by asking three workers to answer them and removing those that a majority of workers answers incorrectly.
- About 17000 training questions and 7000 validation and test questions on
- CLEVR images.

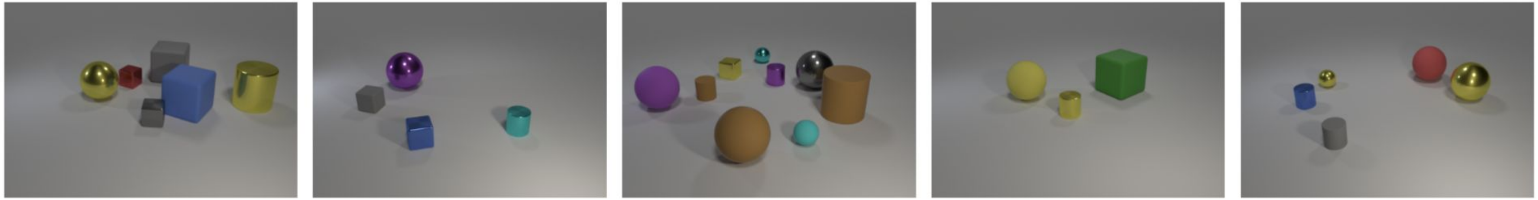
# Experiments

Human-composed questions

| Method | Train CLEVR | Train CLEVR, finetune human |
|---|---|---|
| LSTM | 27.5 | 36.5 |
| CNN+LSTM | 37.7 | 43.2 |
| CNN+LSTM+SA+MLP | 50.4 | 57.6 |
| Ours (18K prog.) | **54.0** | **66.6** |

Table 3. Question answering accuracy on the CLEVR-Humans test set of four models after training on just the CLEVR dataset (**left**) and after finetuning on the CLEVR-Humans dataset (**right**).

# Results



**Q:** *Is there a blue box in the items?* **A:** *yes*

**Predicted Program:**
```
exist
filter_shape[cube]
filter_color[blue]
scene
```

**Predicted Answer:**
✓ *yes*

**Q:** *What shape object is farthest right?* **A:** *cylinder*

**Predicted Program:**
```
query_shape
unique
relate[right]
unique
filter_shape[cylinder]
filter_color[blue]
scene
```

**Predicted Answer:**
✓ *cylinder*

**Q:** *Are all the balls small?* **A:** *no*

**Predicted Program:**
```
equal_size
query_size
unique
filter_shape[sphere]
scene
query_size
unique
filter_shape[sphere]
filter_size[small]
scene
```

**Predicted Answer:**
✓ *no*

**Q:** *Is the green block to the right of the yellow sphere?* **A:** *yes*

**Predicted Program:**
```
exist
filter_shape[cube]
filter_color[green]
relate[right]
unique
filter_shape[sphere]
filter_color[yellow]
scene
```

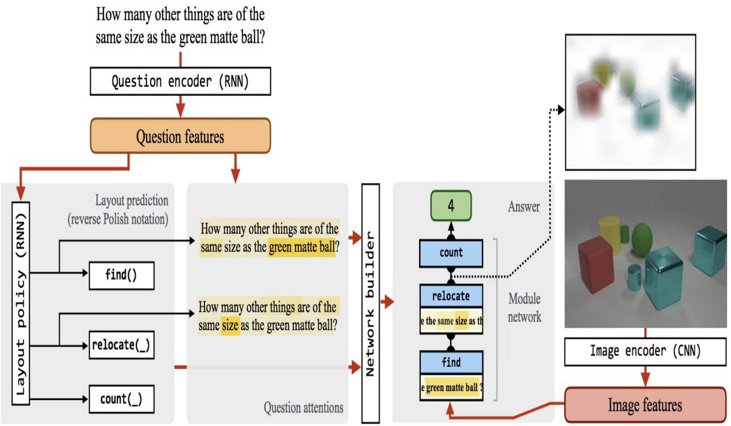
**Predicted Answer:**
✓ *yes*

**Q:** *Two items share a color, a material, and a shape; what is the size of the rightmost of those items?* **A:** *large*

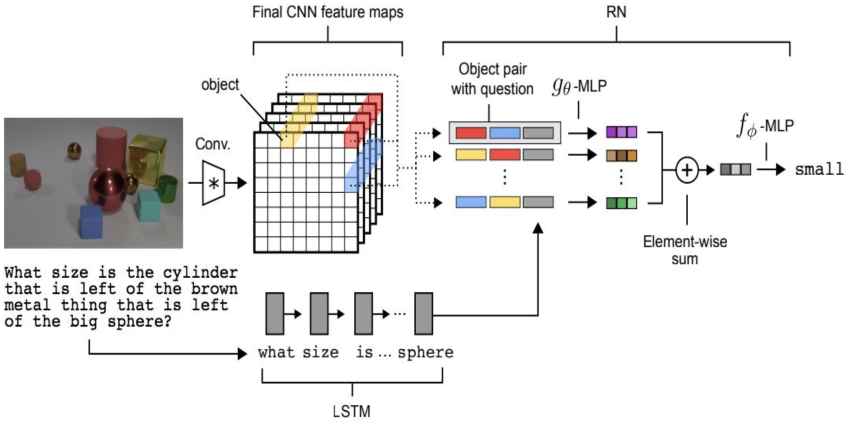
**Predicted Program:**
```
count
filter_shape[cube]
same_material
unique
filter_shape[cylinder]
scene
```

**Predicted Answer:**
✗ *0*

# Other approaches




Andreas et al. ICCV 2017




Santoro et al. arXiv 2017

# Strengths and weaknesses

Strengths

- Novel idea of using compositional reasoning to answer complex questions
- Train program generator on questions using LSTMs
- Training the whole network end to end

Weaknesses

- Not enough results on real images!
- More complex questions may not work properly

# Future works/ possible improvements

Ideas taken from paper

- Adding ternary operations (if/else/then) and loops (for, do) to answer questions like “What color is the object with a unique shape?” .
- Control-flow operators could be incorporated into the framework
- Learning programs with limited supervision

Our ideas

- Using treeRNNs to synthesize programs
- Testing the whole framework on real images

# Conclusion

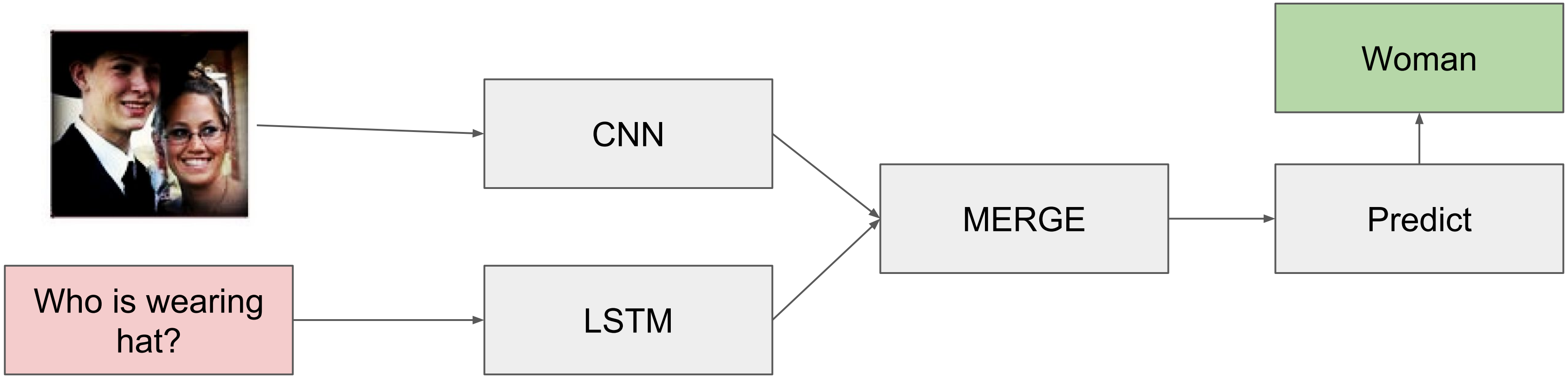
- This method outperforms previous baselines.
- Neural module networks are a more natural way to reproduce reasoning step.
- More flexibility in the composition of the neural module network as modules have generic architectures.

# References


- Inferring and Executing Programs for Visual Reasoning
- CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning
- Talk - https://www.youtube.com/watch?v=3pCLma2FqSk
- Learning to Reason: End-to-End Module Networks for Visual Question Answering
- https://github.com/facebookresearch/clevr-iep
- Deep Compositional Question Answering with Neural Module Networks

# Thanks!

# Visual question answering



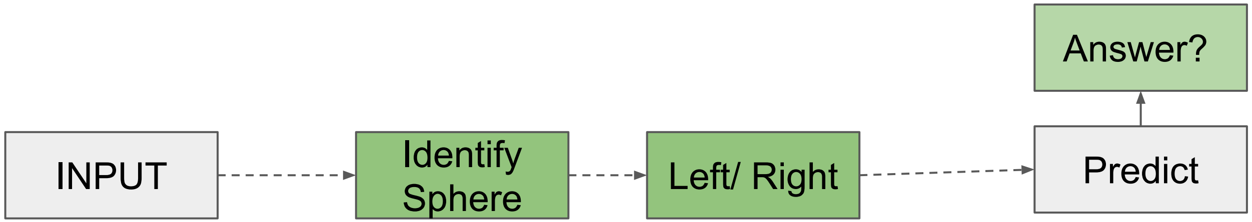

- But does not really understand the question; same answer for
    - who is wearing hat? who is wearing?; wearing?

Antol etal. Vqa: Visual question answering : Proceedings of the IEEE International Conference on Computer Vision

# Standard VQA?

**Q**: How many spheres are the **left** of the big sphere and the same color as the small rubber cylinder?

**Q**: How many spheres are the **right** of the big sphere and the same color as the small rubber cylinder?




**Decompose the network!!**