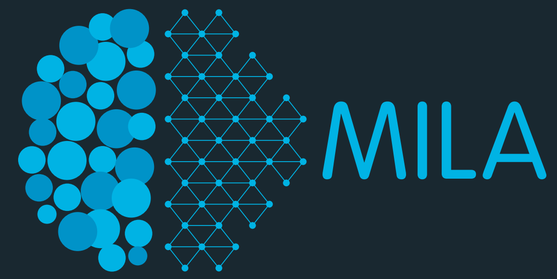


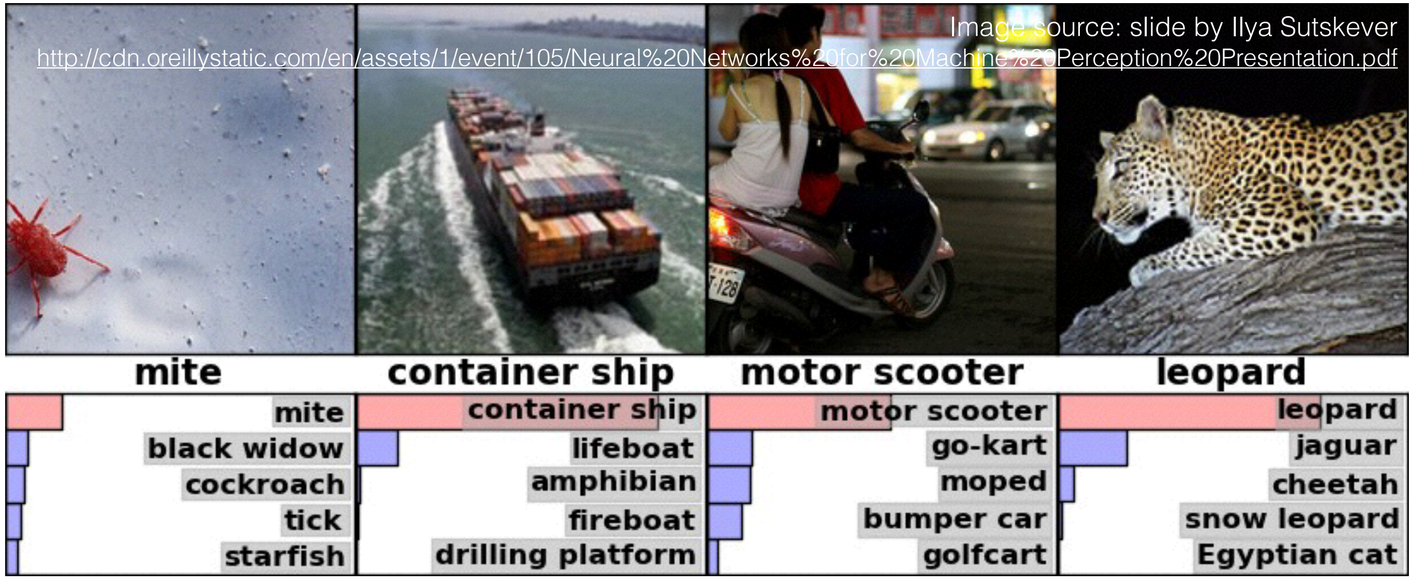
Visual Reasoning via Feature-wise Linear Modulation

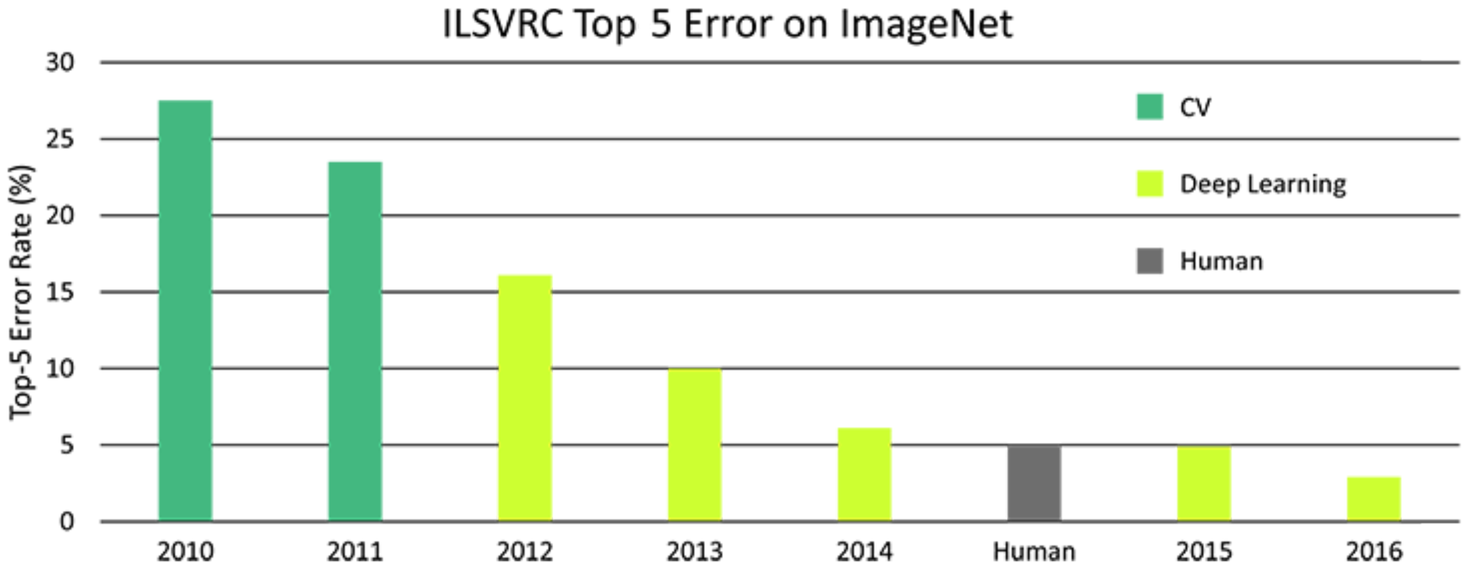
Aaron Courville
MILA
Université de Montréal

Deep Learning for Vision

Rapid progress in recent years on basic object recognition.

1. Convolutional Neural Networks.
2. Accessibility of large datasets.
3. Increase in computational capacity (e.g. GPUs).
4. Real advances in the models and learning algorithms.



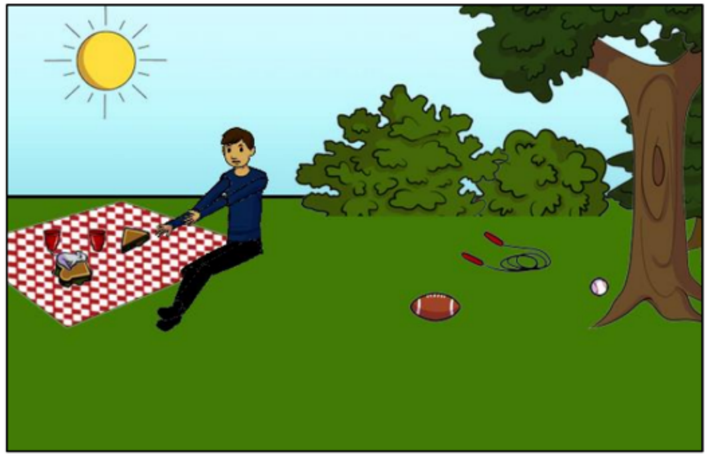


Visual Question Answering

- VQA for vision
  - Allows for a fine-grain semantic probe of a visual scene.
- VQA for language
  - Supports the grounding of language.
  - word meaning beyond statistical association.

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

Agrawal et al. VQA: *Visual Question Answering*. ICCV 2015.

GuessWhat?!

Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin,
Hugo Larochelle, Aaron Courville, CVPR 2017

MILA

- Visual task-oriented dialogue game:
  - Data collection: humans players incentivized to cooperate.
  - Clear objective / evaluation for both the human players and models.

- Two Players: **Oracle, Questionner** both see an image.
  - **Oracle**: sees target object and answers Questionners questions **(Yes / No)**
  - **Questionner**: asks natural language questions of Oracle, tries to guess the target object.

Is it human?	*Yes*
Are they five humans visible?	*Yes*
Is it leftmost?	*No*
Is it in the middle?	*Yes*
Is it the third one from the left?	*Yes*

GuessWhat?!: 2 player game

Questionner

Oracle

GuessWhat?!: 2 player game

Questionner

Oracle

GuessWhat?!: 2 player game

MILA

Questionner	Oracle
Is it a bird?	Yes
Is it left side?	No
Starting from the right, is it the 1st one?	No
One of the two birds?	Yes
Is it the upper one?	Yes

GuessWhat?!: 2 player game

Questionner	Oracle
Is it a person?	Yes
Are they a player?	No
Are they in the stands?	Yes
Are they wearing white?	No
Are they sitting?	Yes
Are they in the front?	No
Are they next to the orange siding?	Yes
Is it the person on the left hand side - wearing a white scarf?	No
Are they wearing dark clothing?	Yes
Are their arms visible?	No

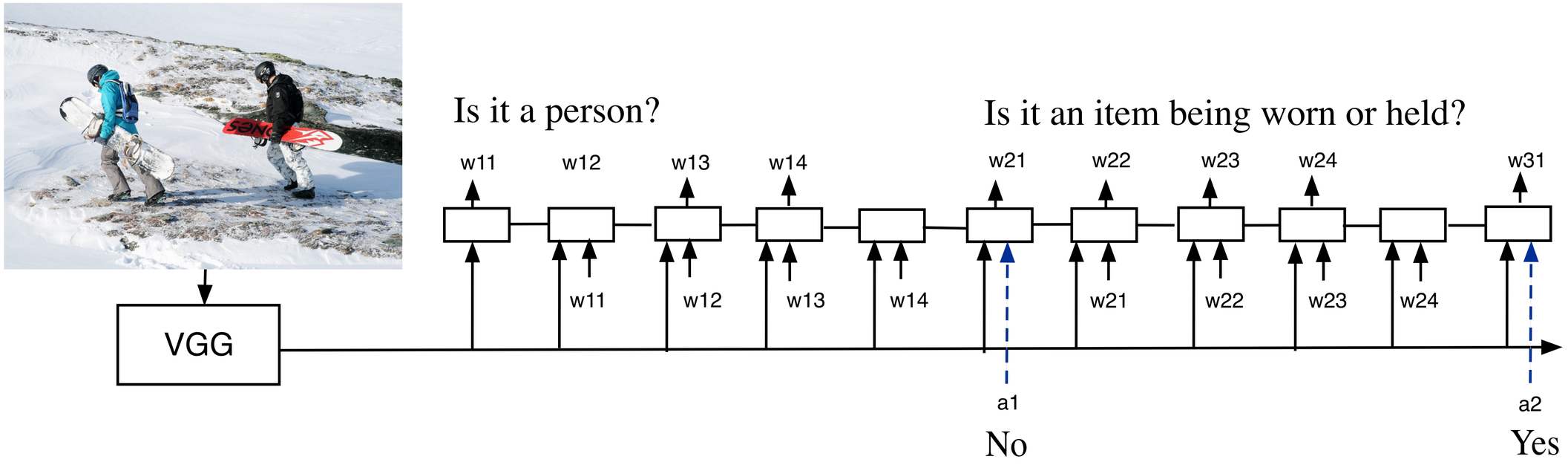
GuessWhat?!: 2 asymmetric agents

- Oracle: supervised learning task
  - Inputs: image, questions, object category/location.
  - Output: Yes / No answer.
- Questionner: reinforcement learning task
  - Inputs: image, previous Questions / Answers.
  - Output: current question.

Is it a person?	*No*
Is it a cat?	*Yes*
Is it in the hands of these girls?	*No*
The cat in the right side of the image?	*Yes*

Questionner: A Reinforcement Learning Task

- Consider the Oracle as fixed (i.e. part of the environment.)
- Action space: words in the queries.
- Reward: **1** if the Guesser guesses the correct object, **0** if not.



End-to-end optimization of goal-driven and visually grounded dialogue systems, IJCAI 2017
Florian Strub, Harm de Vries, Jérémie Mary, Bilal Piot, Aaron Courville, Olivier Pietquin

Questionner: A Reinforcement Learning Task

Baselines are trained to imitate human questions

RL directly learns to ask questions that leads to accurate guessing.

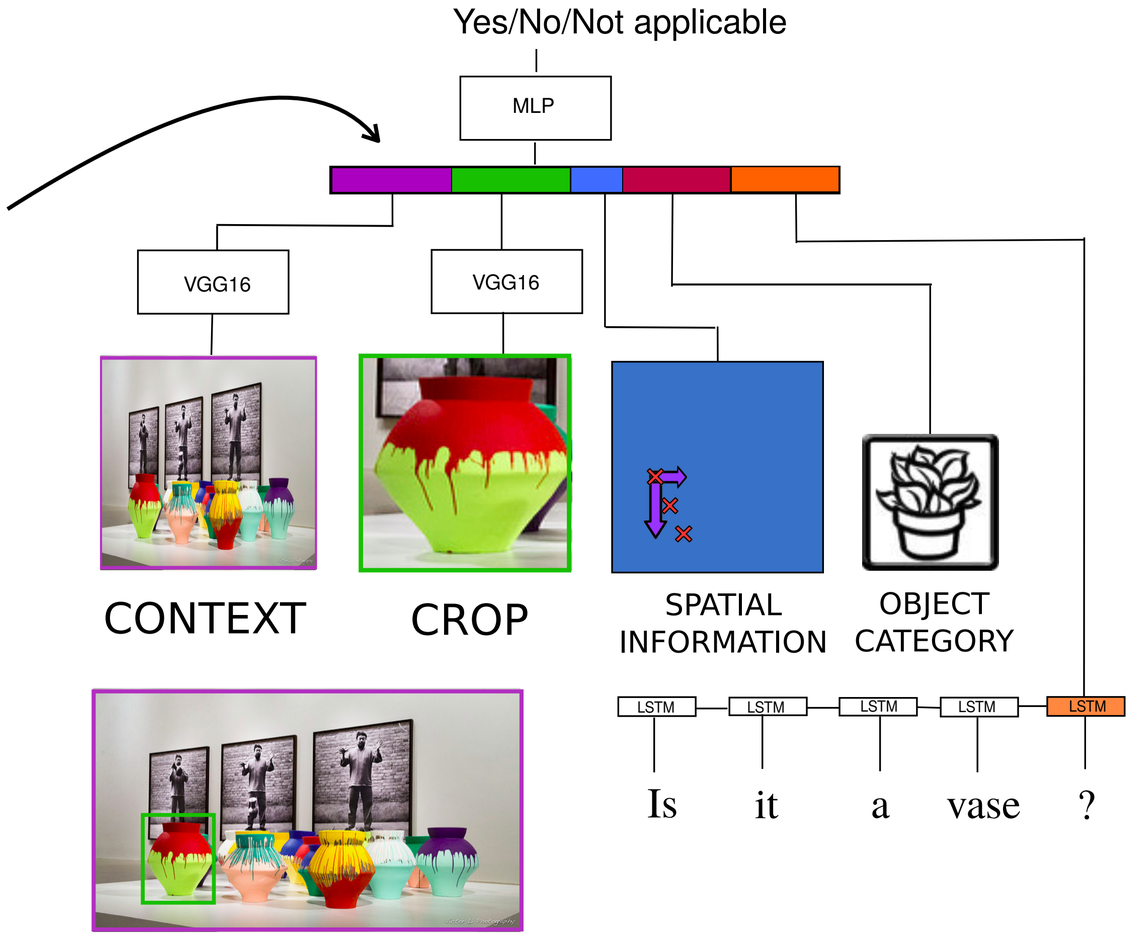
		New Objects	New Pictures
Baseline	Sampling	46.4% ± 0.2	45.0% ± 0.1
	Greedy	48.2% ± 0.1	46.9%
	BSearch	53.4% ± 0.0	53.0%
REINFORCE	Sampling	**63.2% ± 0.3**	**62.0% ± 0.2**
	Greedy	58.6% ± 0.0	57.5%
	BSearch	54.3% ± 0.1	53.2%

- RL improves the success rate of dialogues (% above are relative to the human performance).
- Improvement is due to RL learning to avoid difficult questions for the Oracle.
- Oracle Q/A accuracy limits the ability of RL to improve performance.

End-to-end optimization of goal-driven and visually grounded dialogue systems, IJCAI 2017
Florian Strub, Harm de Vries, Jérémie Mary, Bilal Piot, Aaron Courville, Olivier Pietquin

Can we improving the Oracle?

- Baseline Oracle:
  - Concatenation of the image, question and object features into a classification MLP
  - We can do better.



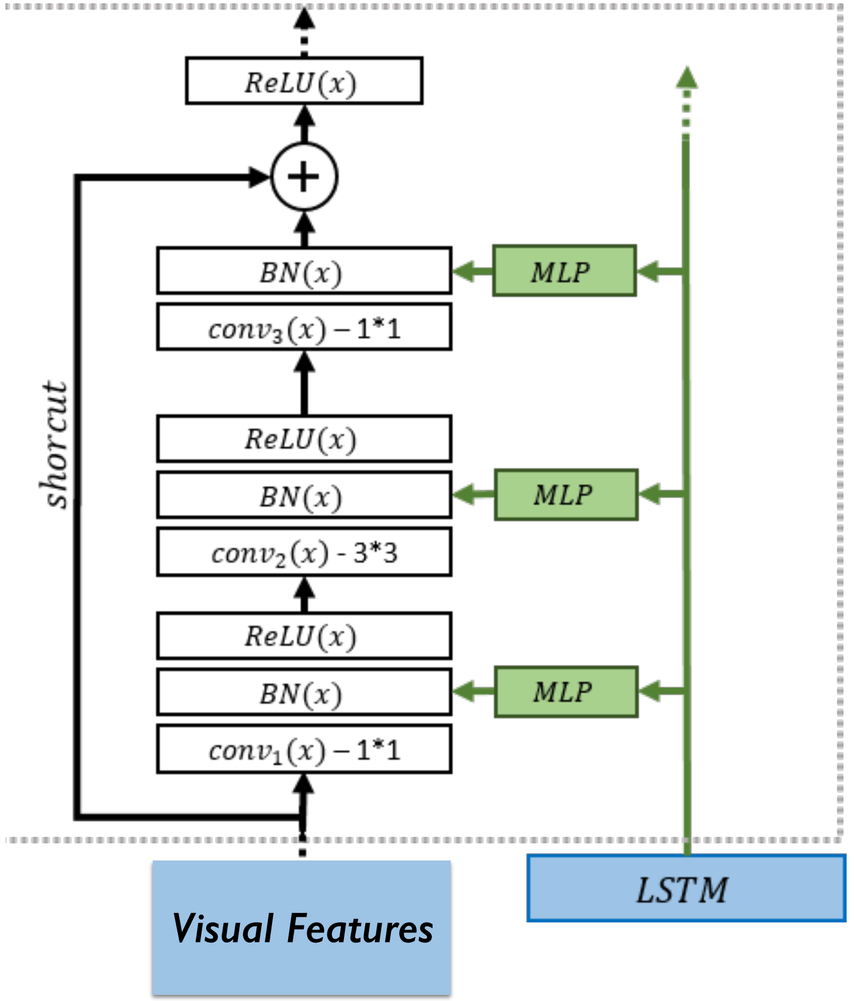
Modulating early visual processing by language. NIPS 2017

Harm de Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, Aaron Courville

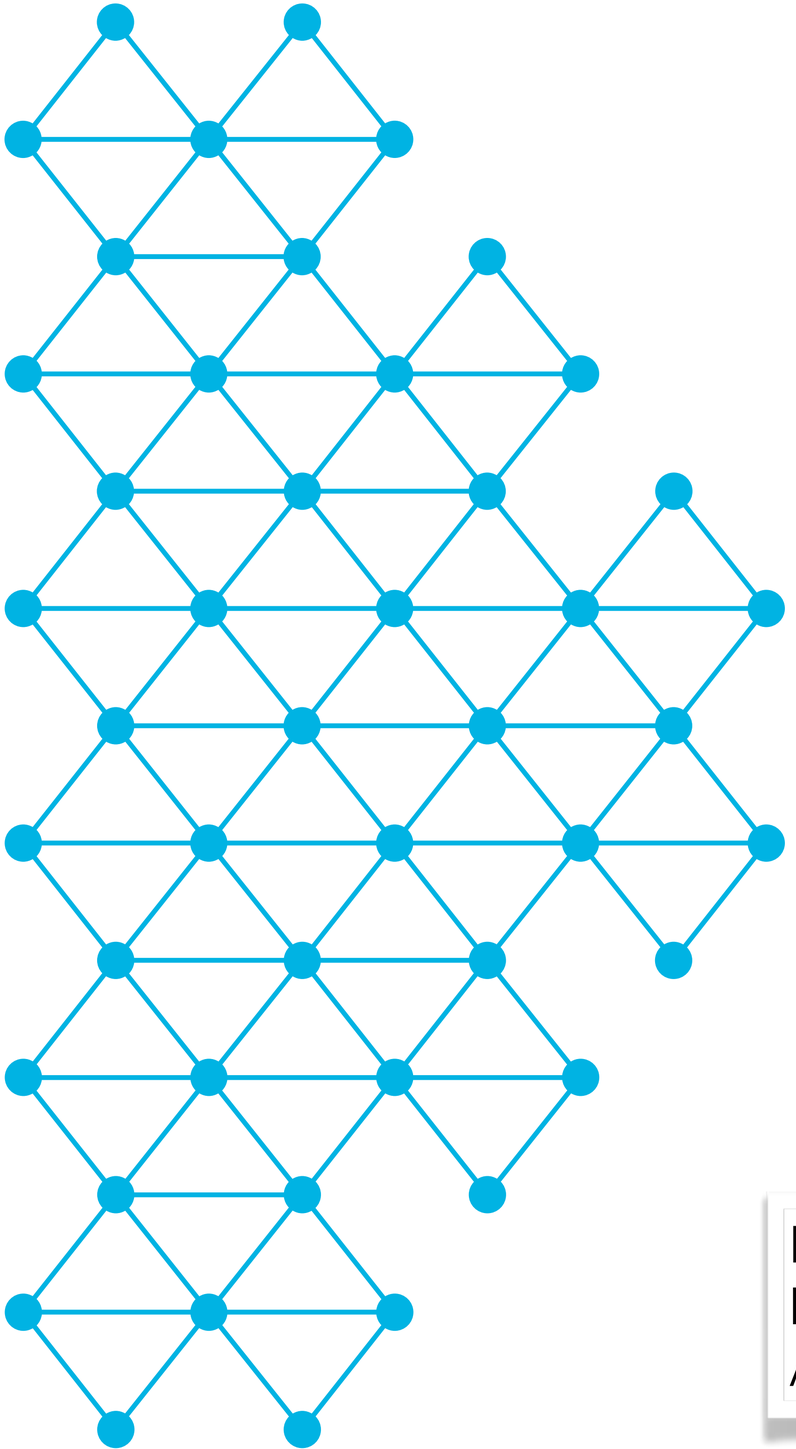
Can we improving the Oracle?

- Modulate early visual processing by Conditioned Batch Normalization.
  - Inspiration: Dumoulin et al. ICLR 2017.
    A Learned Representation For Artistic Style.

- *We will explore this form of integration of questions into an image pipeline in another, relate context:* ***Visual Reasoning****.*



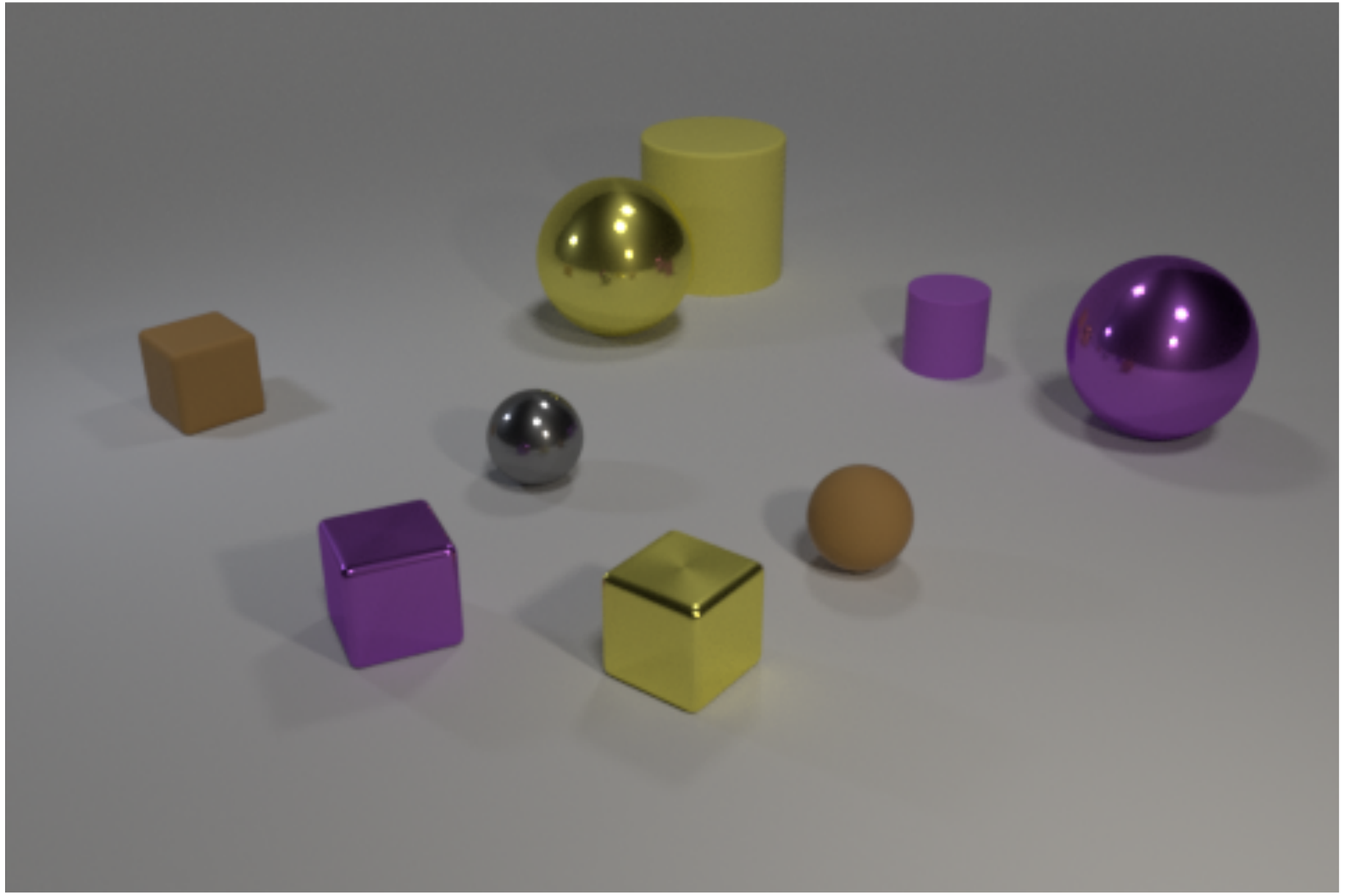
Modulating early visual processing by language. NIPS 2017
Harm de Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, Aaron Courville



Conditioned Normalization for Visual Reasoning

FiLM: Visual Reasoning with a General Conditioning Layer.
Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, Aaron Courville
AAAI 2017 (submission) and currently on ArXiv

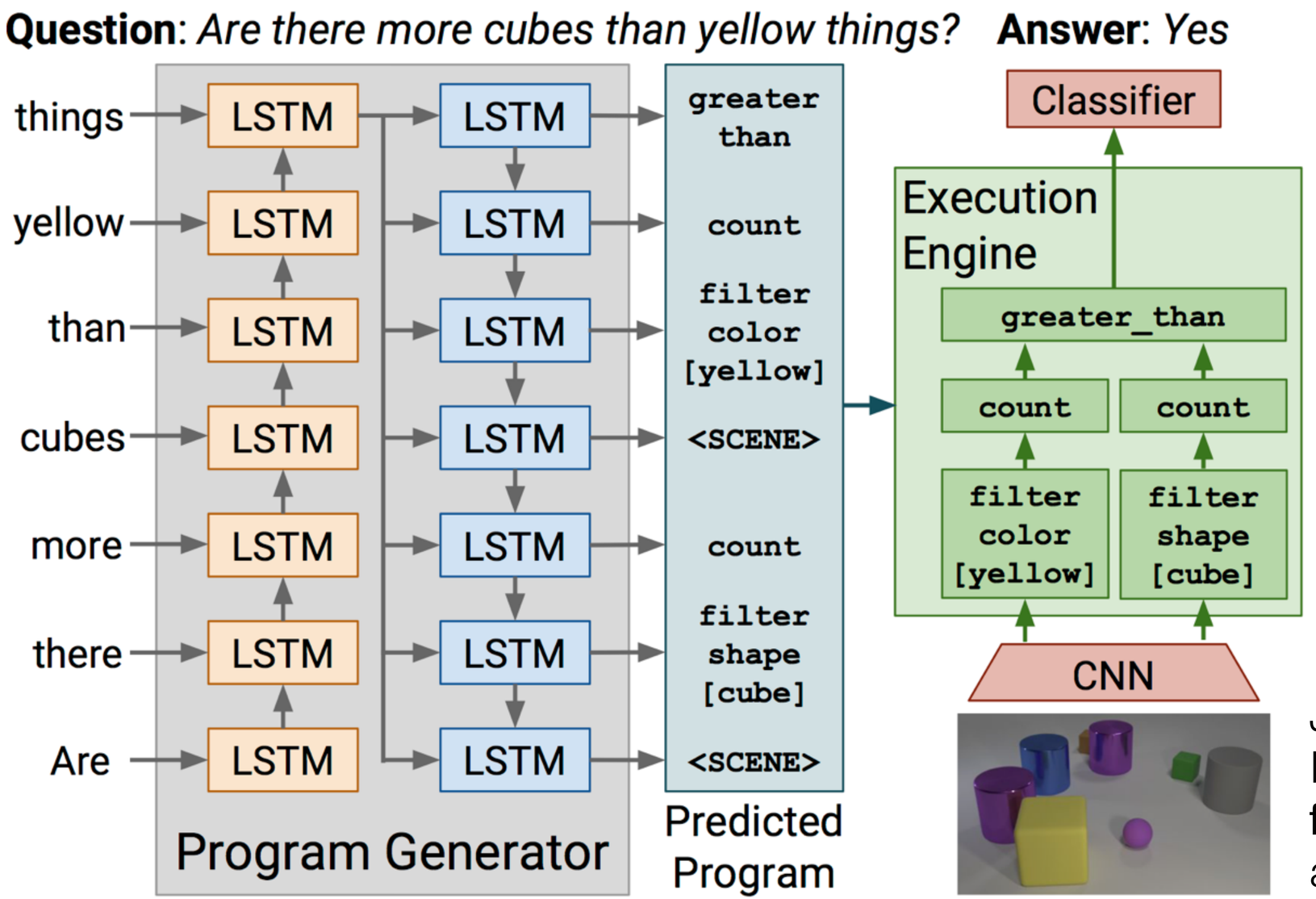
Visual Reasoning: CLEVR Dataset



Q: What number of cylinders are small purple things or yellow rubber things?

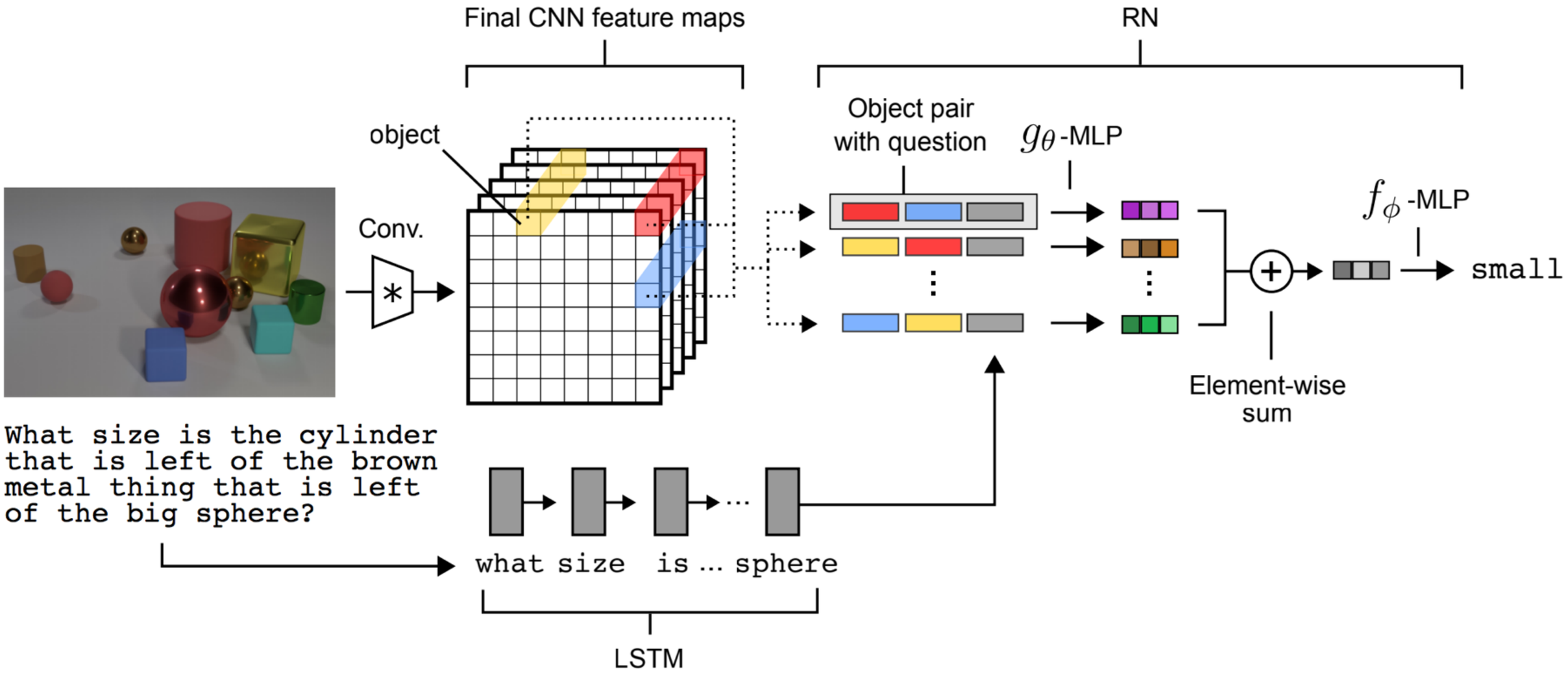
A: 2

Previous Work: Program Generation



Johnson et al.
Inferring and Executing Programs for Visual Reasoning.
arXiv 2017.

Previous Work: Relational Networks



Santoro et al. **A simple neural network module for relational reasoning.** arXiv 2017.

Previous Work: CLEVR Accuracy

Model	Overall
Human (Johnson et al. 2017b)	92.6
Q-type baseline (Johnson et al. 2017b)	41.8
LSTM (Johnson et al. 2017b)	46.8
CNN+LSTM (Johnson et al. 2017b)	52.3
CNN+LSTM+SA (Santoro et al. 2017)	76.6
N2NMN* (Hu et al. 2017)	83.7
PG+EE (9K prog.)* (Johnson et al. 2017b)	88.6
PG+EE (700K prog.)* (Johnson et al. 2017b)	96.9
CNN+LSTM+RN†‡ (Santoro et al. 2017)	95.5

— Program Generation

— Relational Networks

Lesson(?): achieving high accuracy

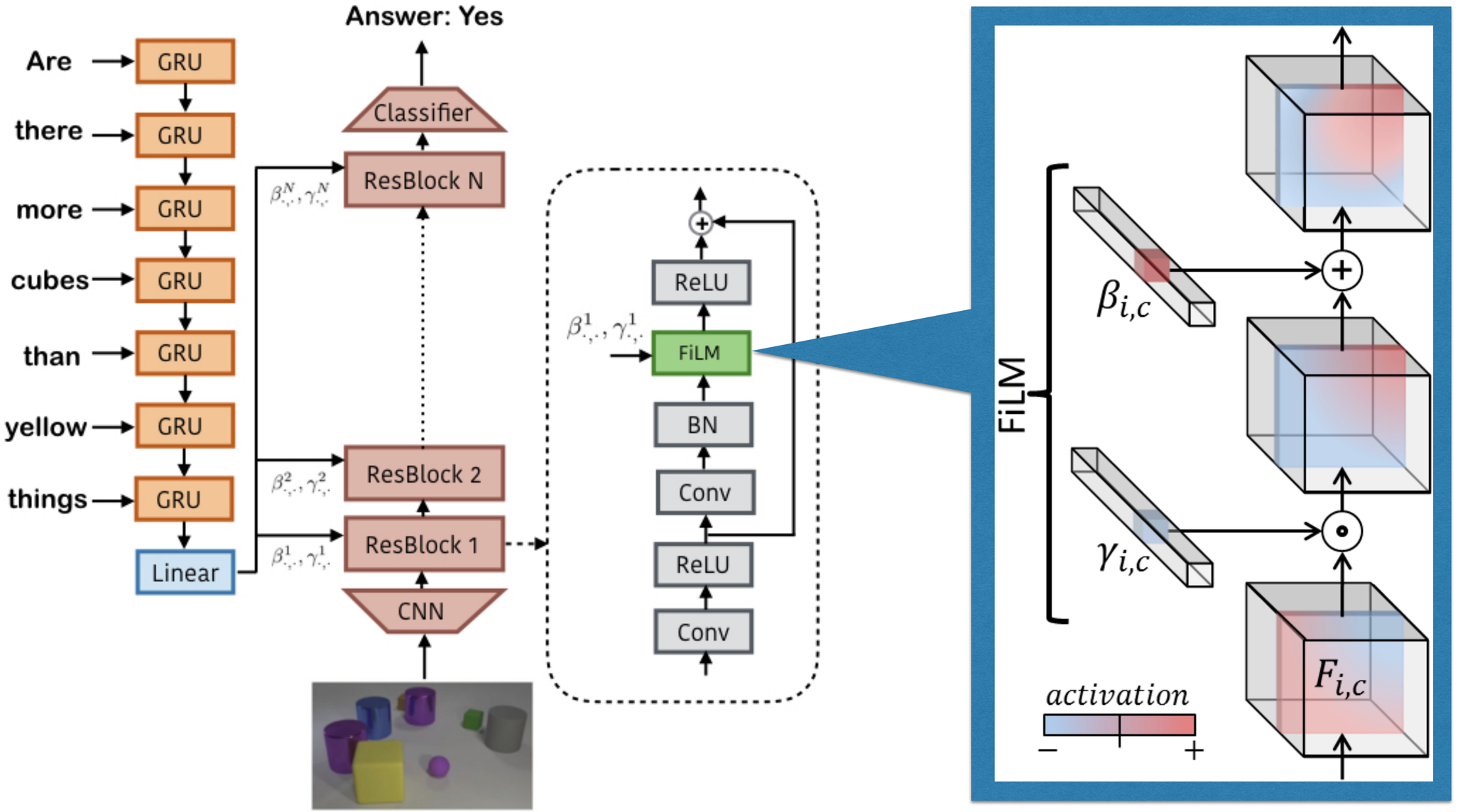
- requires the use of extra (program) data.
- or task-oriented model design.

∗ program label use

† data augmentation

‡ training from pixels

FiLM-Based Visual Reasoning



FiLM: CLEVR Accuracy

Model	Overall
Human (Johnson et al. 2017b)	92.6
Q-type baseline (Johnson et al. 2017b)	41.8
LSTM (Johnson et al. 2017b)	46.8
CNN+LSTM (Johnson et al. 2017b)	52.3
CNN+LSTM+SA (Santoro et al. 2017)	76.6
N2NMN* (Hu et al. 2017)	83.7
PG+EE (9K prog.)* (Johnson et al. 2017b)	88.6
PG+EE (700K prog.)* (Johnson et al. 2017b)	96.9
CNN+LSTM+RN†‡ (Santoro et al. 2017)	95.5
CNN+GRU+FiLM	**97.7**
CNN+GRU+FiLM‡	97.6

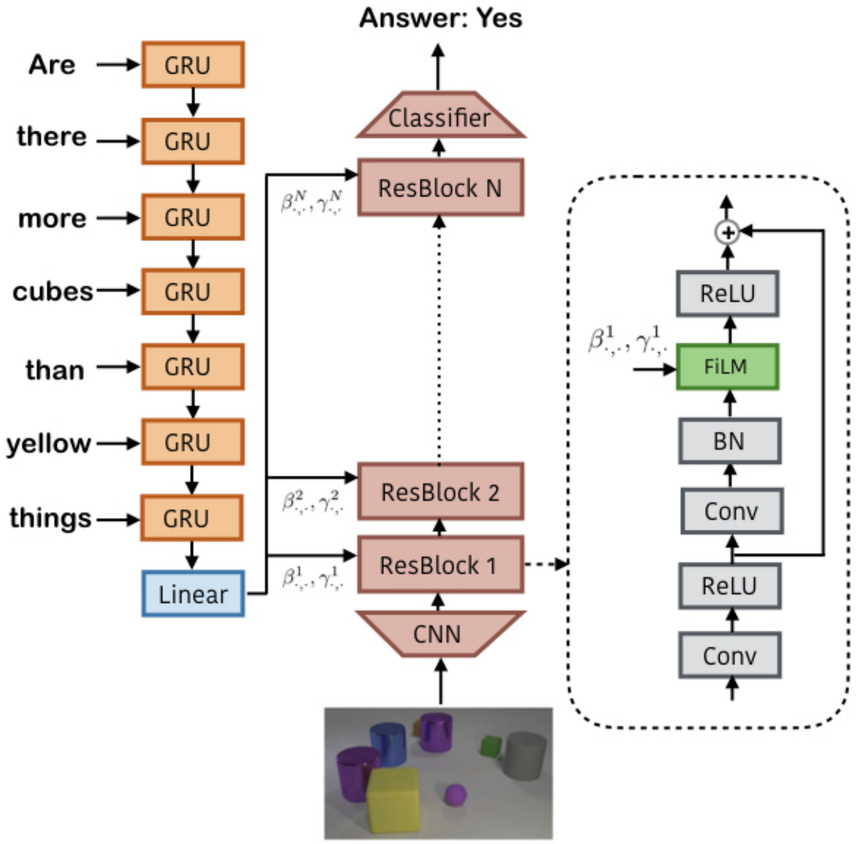
* program label use

† data augmentation

‡ training from pixels

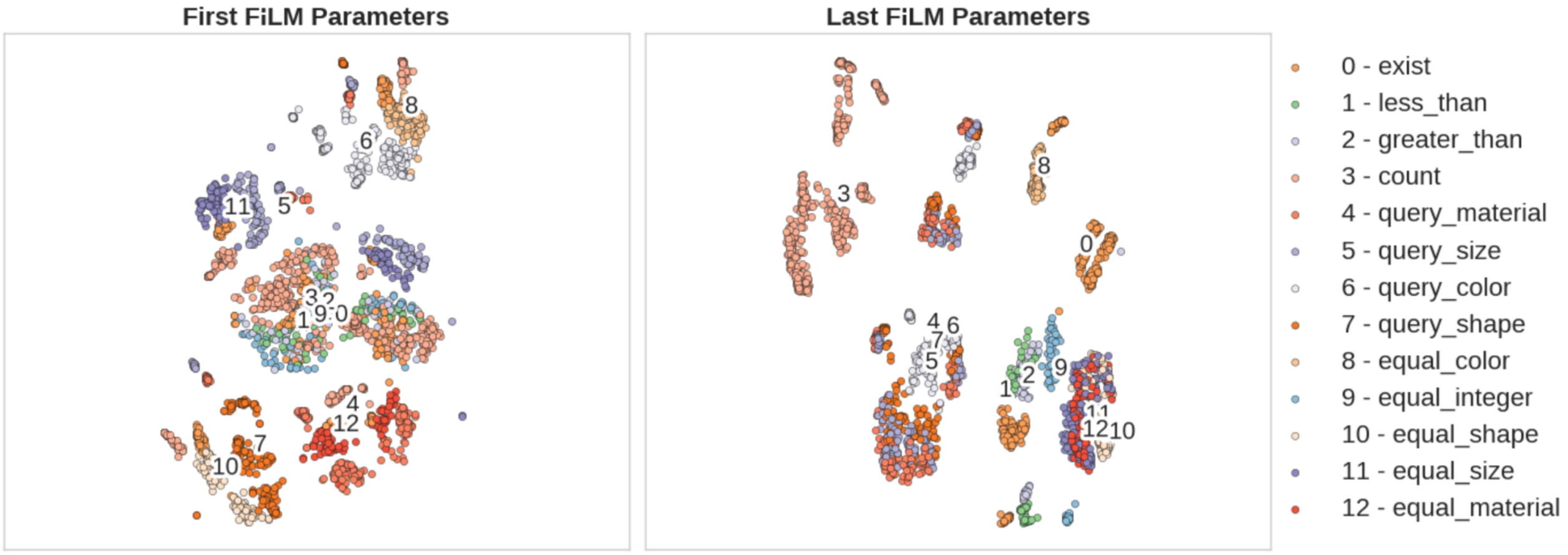
FiLM achieves SOTA accuracy

- Despite no use of program data.
- Without strongly task-oriented model design.



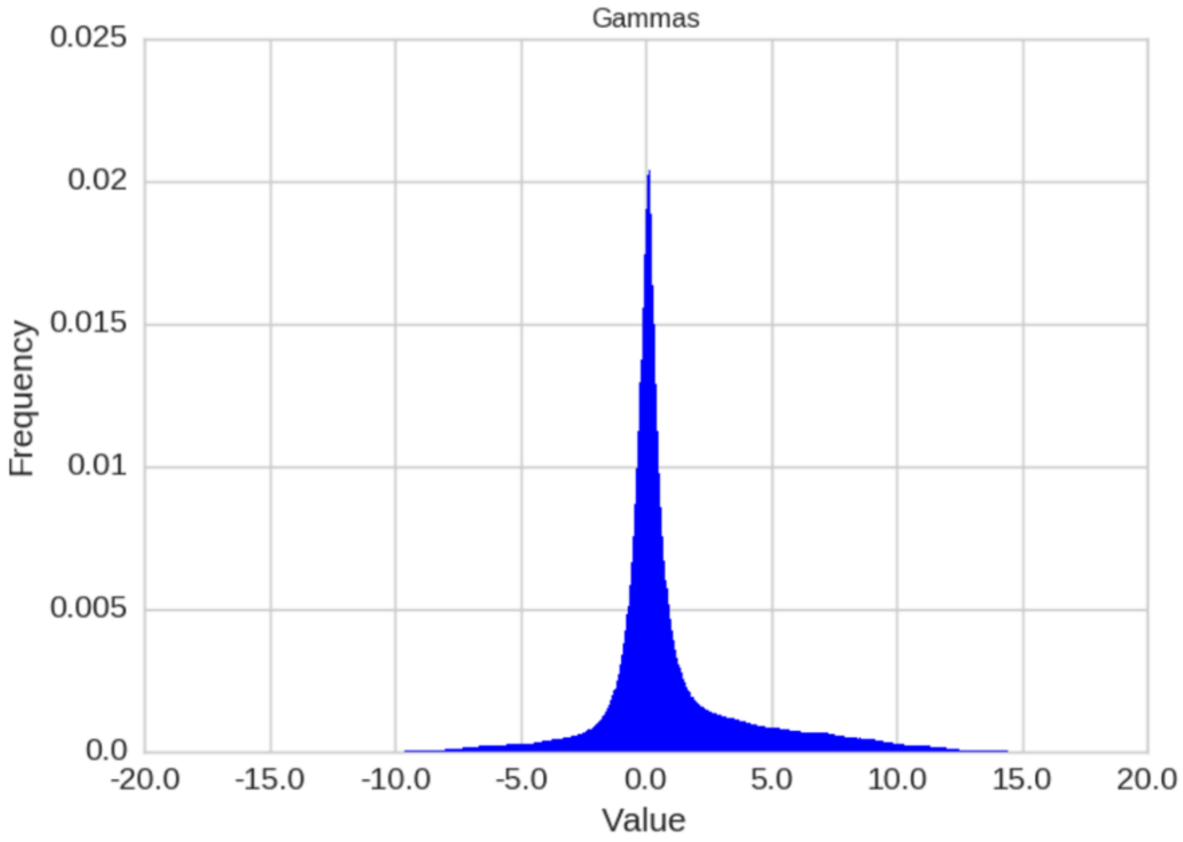
How does FiLM reason?

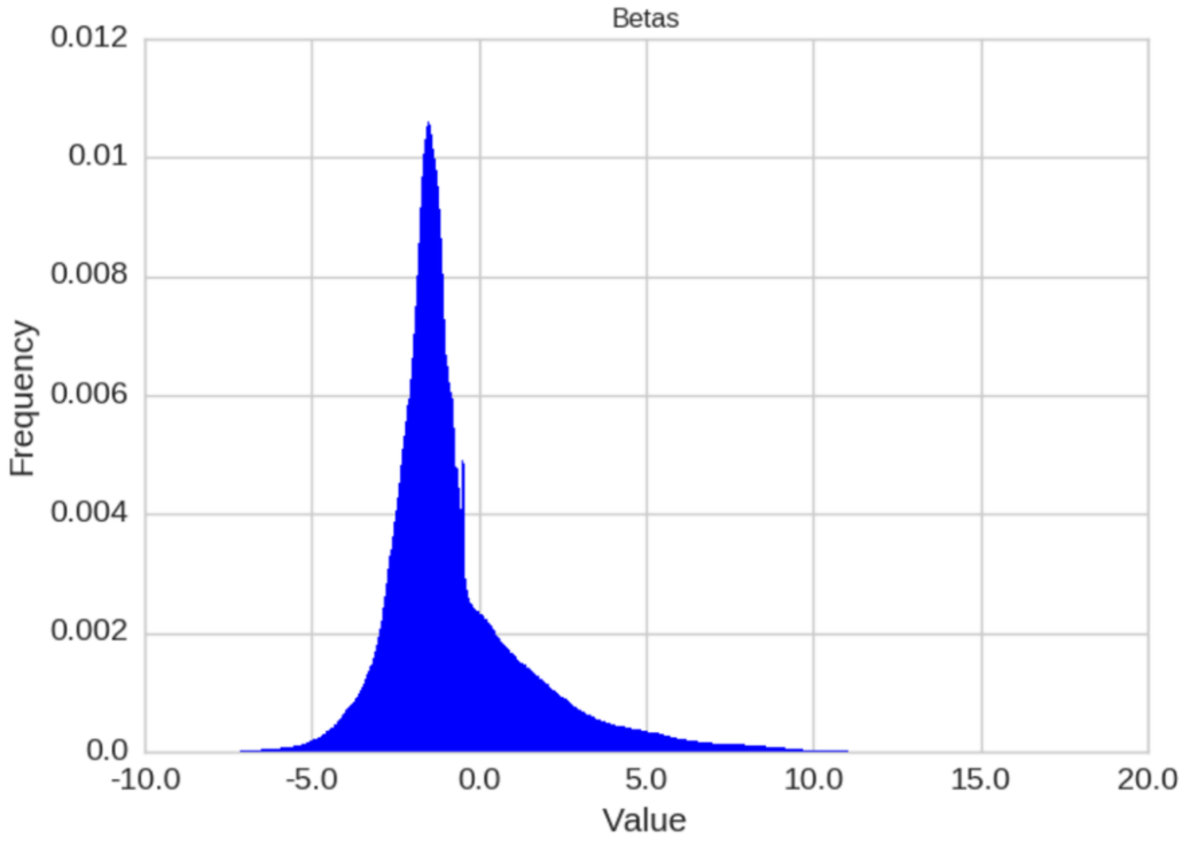
TSNE projections of FiLM's shift and scaling values



How does FiLM reason?

Histograms of $\gamma_{i,c}$ and $\beta_{i,c}$ values





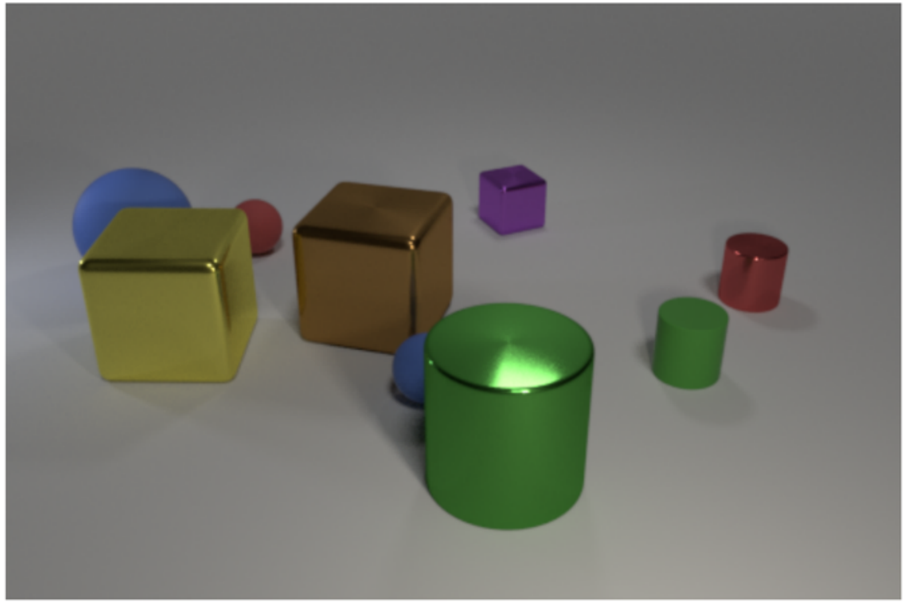
Generalizing to Human Questions

MILA

CLEVR-Humans Task:

- Human-posed questions
  - Free-form
  - More complex
  - New vocab (underlined)
- Small: 18K training samples

Model	Train CLEVR	Train CLEVR, fine-tune human
LSTM	27.5	36.5
CNN+LSTM	37.7	43.2
CNN+LSTM+SA+MLP	50.4	57.6
PG+EE (18K prog.)	54.0	66.6
CNN+GRU+FiLM	**56.6**	**75.9**



Q: *Which shape objects are partially obscured from view?* **A:** *Sphere*

Conclusions

- VQA, Visual Dialogue and Visual Reasoning tasks offer a means to:

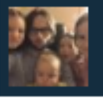
  a) Explore detailed semantics in a visual scene.

  b) Ground language in the physical word.

- Details matter: the right architectural features can make a big difference to performance.

- FiLM and related conditional scaling and shifting operations are effective means of integrating language into a visual task.

GUESSWHAT?!

Home People Download Explore Play the game 

MILA

Game instructions

This is a two-player game, yourself and a partner. You will be (randomly) assigned to play one of the two roles:

Questioner
Find the object

- You will be shown an image of a scene with multiple objects.
- One of the objects will be assigned as the target (but not visible to you).
- Your job is to locate that object by asking yes or no questions.
- You click on the GuessWhat! button once you are certain which object it is.
- All object segmentations are then shown in the image, and you click on the correct object.

Oracle
Answer the questions

- You will be shown an image of a scene with multiple objects.
- One of the objects will be assigned as the target.
- Your partner will ask yes/no questions to locate this object.
- Your job is to answer their questions correctly.

Ready to play?

Play with the AI Oracle!

Start a game »

Play with AI »

https://guesswhat.ai