2A
Deep Compositional Question Answering With Neural Module Networks

GHAZAL SAHEBZAMANI, DELARAM BEHNAMI, PARDISS DANAEI, GOLARA JAVADI

19 MARCH 2019

Introduction

- Visual question answering based on neural module networks (NMNs)
- Understanding of both visual scenes and natural language

What color is the necktie?

Yellow

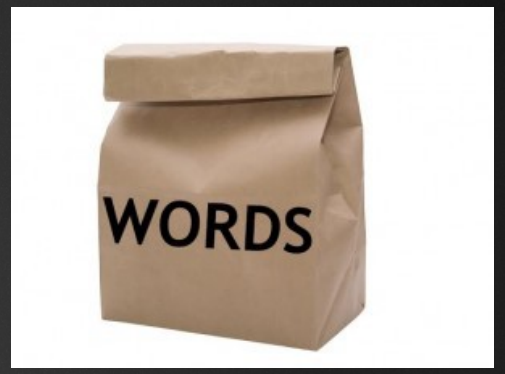
Acknowledgement

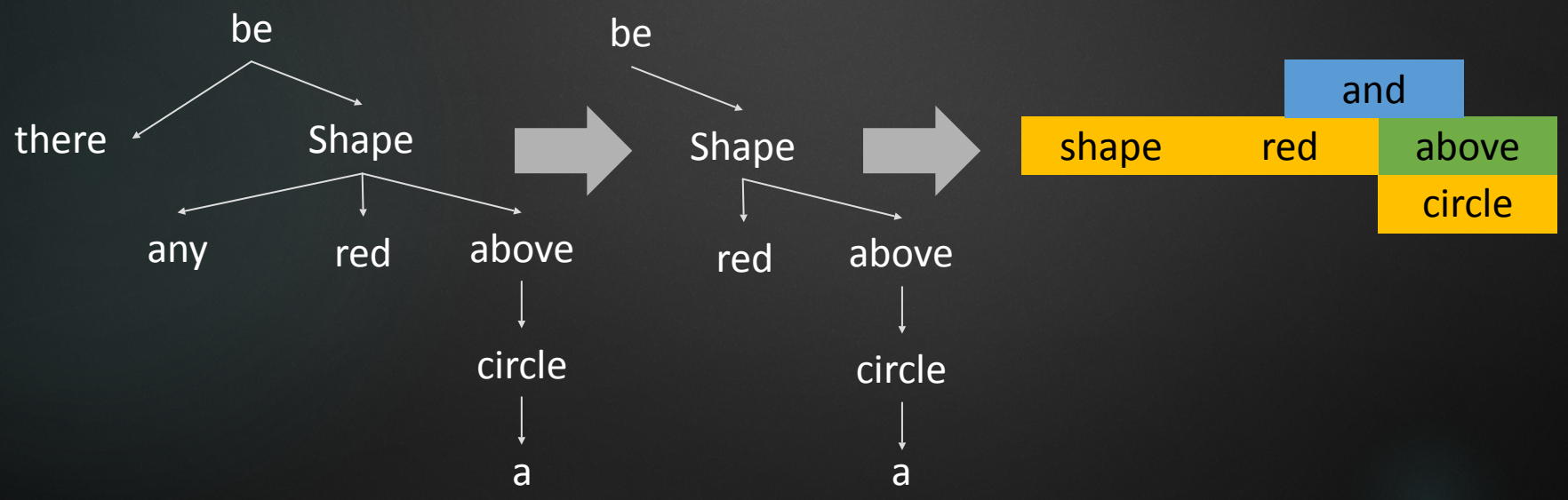
- PAPER 1
  - Andreas J, Rohrbach M, Darrell T, Klein D. Neural module networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016 (pp. 39-48).
  - Andreas J, Rohrbach M, Darrell T, Klein D. Learning to compose neural networks for question answering. 2016. PowerPoint Presentation.
- PAPER 2
  - Wu Q, Teney D, Wang P, Shen C, Dick A, van den Hengel A. Visual question answering: A survey of methods and datasets. Computer Vision and Image Understanding. 2017 Oct 1;163:21-40.
- PAPER 3
  - Hu R, Andreas J, Darrell T, Saenko K. Explainable neural computation via stack neural module networks. In Proceedings of the European Conference on Computer Vision (ECCV) 2018 (pp. 53-69).

NMN: Assemble a network on the fly from a collection of jointly-learned modulesAnalyze question with semantic parserDetermine computational units based on the analysisGround the

orks

- Neural networks representation
  - Represent question as bags of words
  - Represent question using a recurrent neural network
  - Train a simple classifier on the encoded question and image
- Semantic Parsers





NMN: Assemble a network on the fly from a collection of jointly-learned modules
Common to initialize systems for vision tasks with a prefix of a network trained

proach

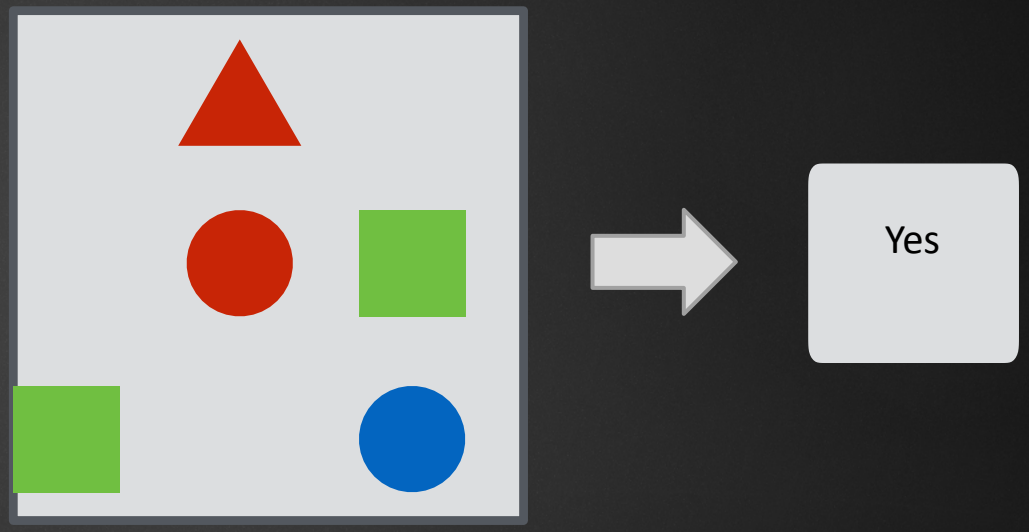
- NMN: Assemble a network on the fly from a collection of jointly-learned modules
  - No single “best network” for all tasks
  - Question answering is a highly-multitask learning setting

Ground the question on the image
Represent images with visual features and attention
Messages passed between modules may be raw image features, attentions,

proach

- Ground the question on the image
  - Neural nets learn lexical groundings

Is there a red shape above a circle?



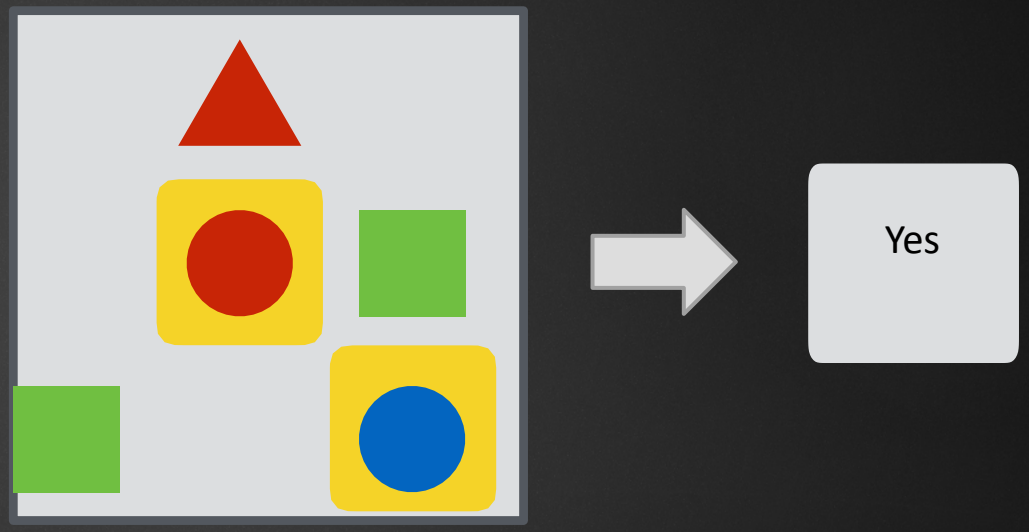
- Use a recurrent network (LSTM) to read the question for final answer

Ground the question on the image
Represent images with visual features and attention
Messages passed between modules may be raw image features, attentions,

proach

- Ground the question on the image
  - Neural nets learn lexical groundings

Is there a red shape above a circle?



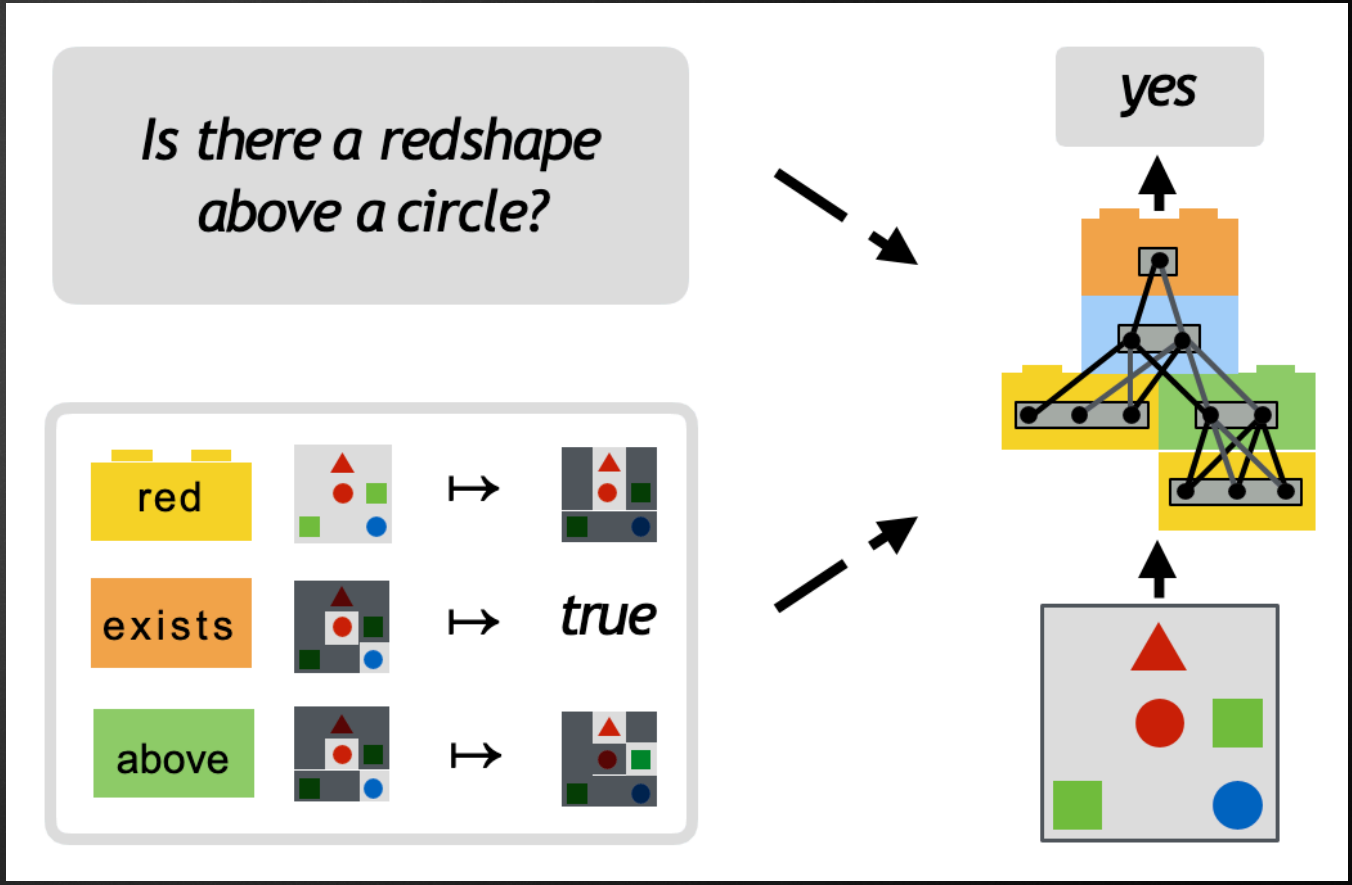
- Use a recurrent network (LSTM) to read the question for final answer

each module has a set of parameters, and a Network layout predictor

Instantiate network based on network predictor

Neural Module Networks (NMNs)

- Model is a collection of modules
- Training datum
  - Natural language question
  - Image
  - Answer
- Classifier



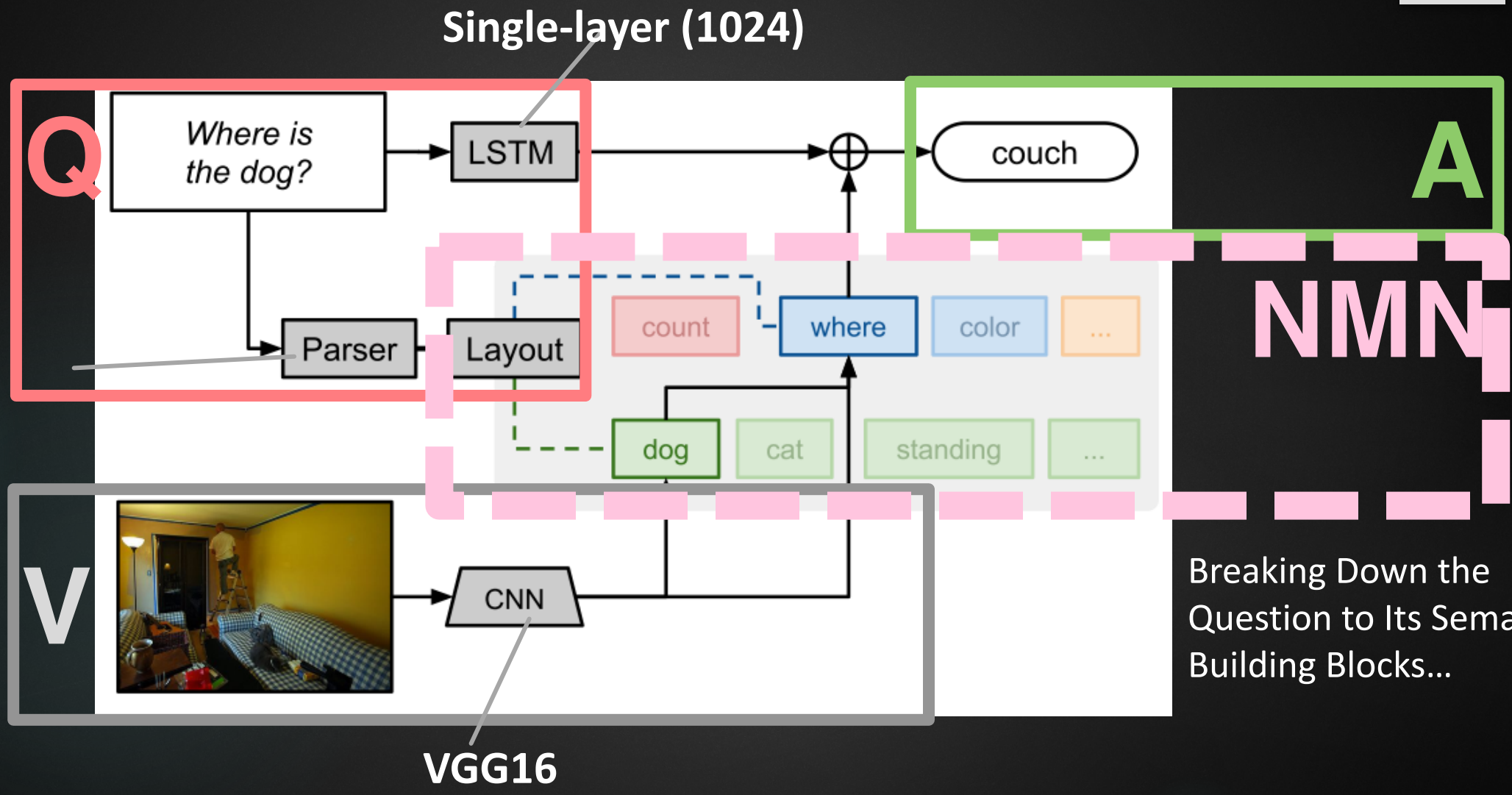
VQA System

V

Q *Is there a red shape above a circle?*

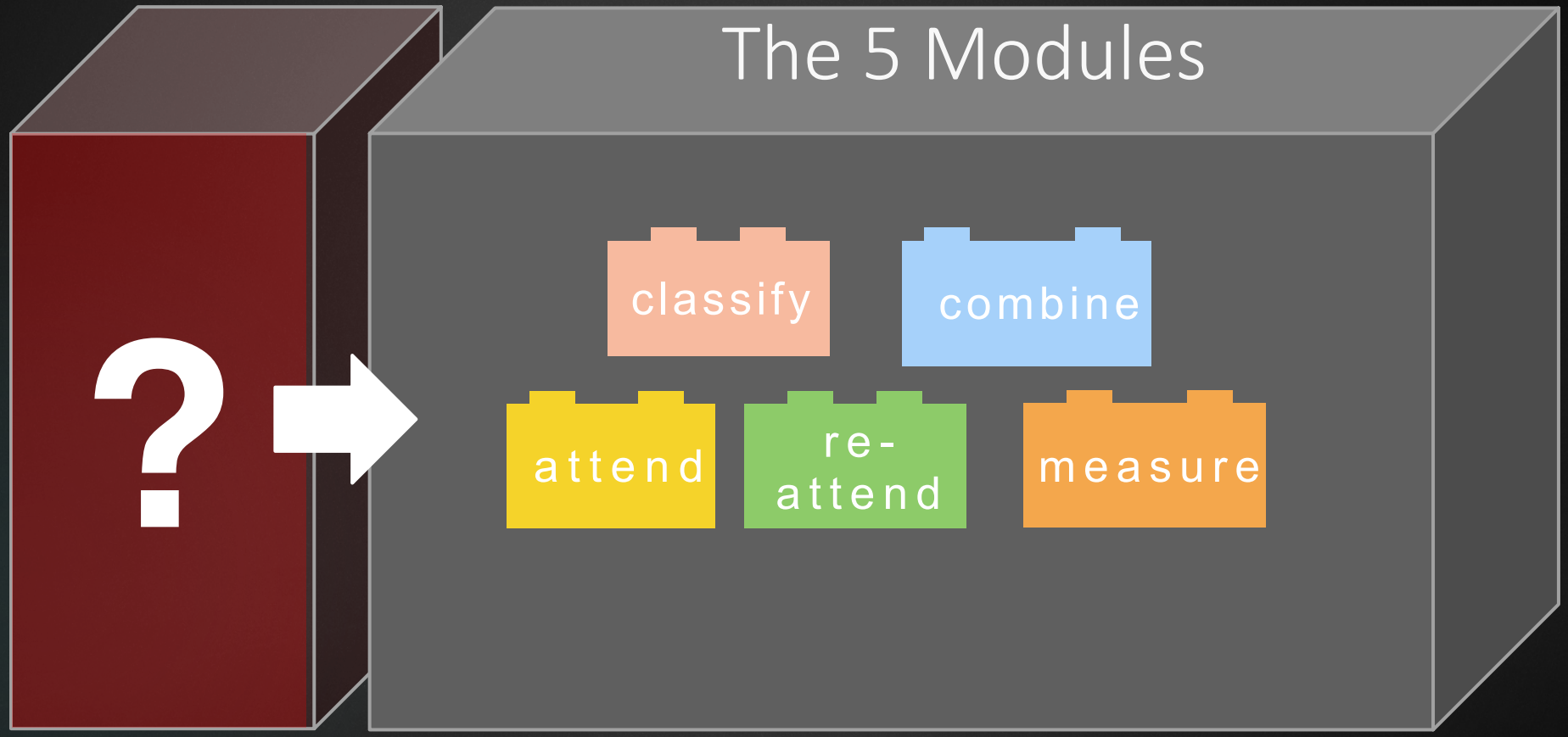
A *Yes.*

Compositional VQA with NMNs

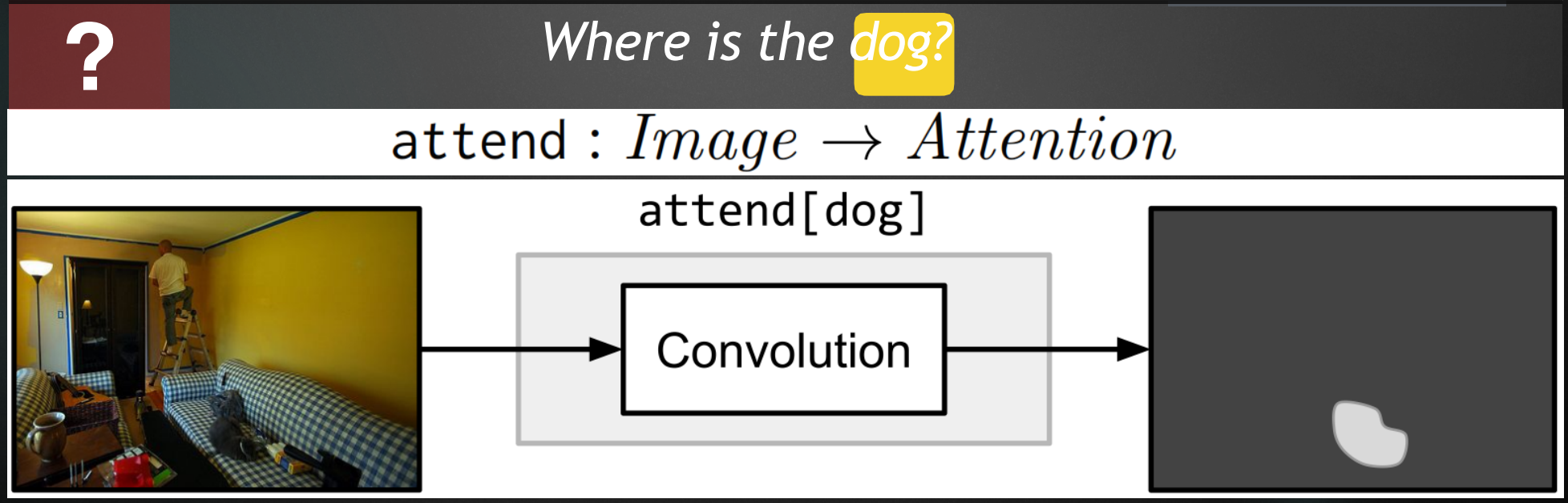


Breaking Down the Question to Its Semantic Building Blocks…

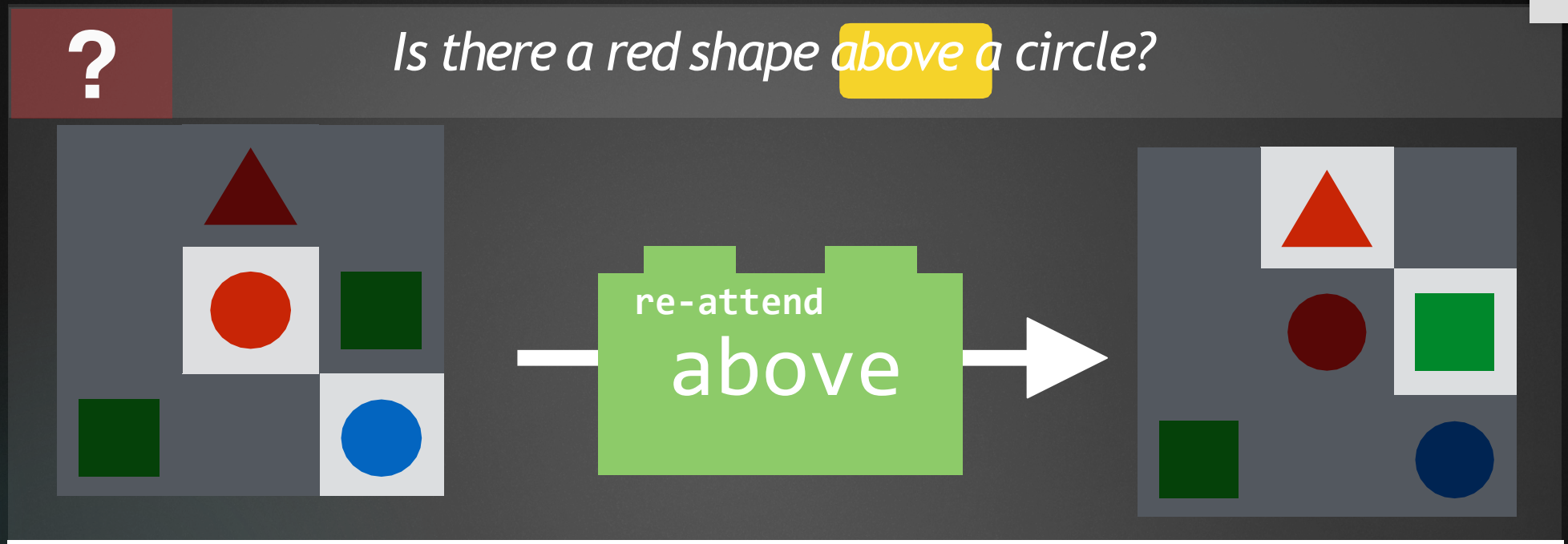
What Are the Building Blocks of Questions in VQA?



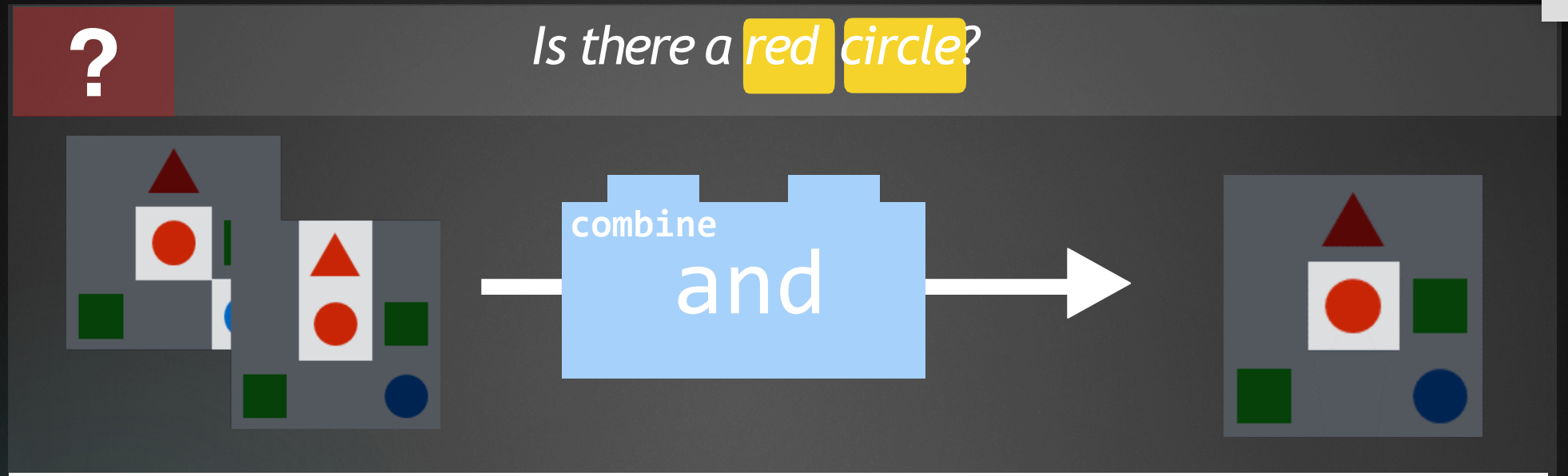
1. The **attend** Module for Attention to Predicates



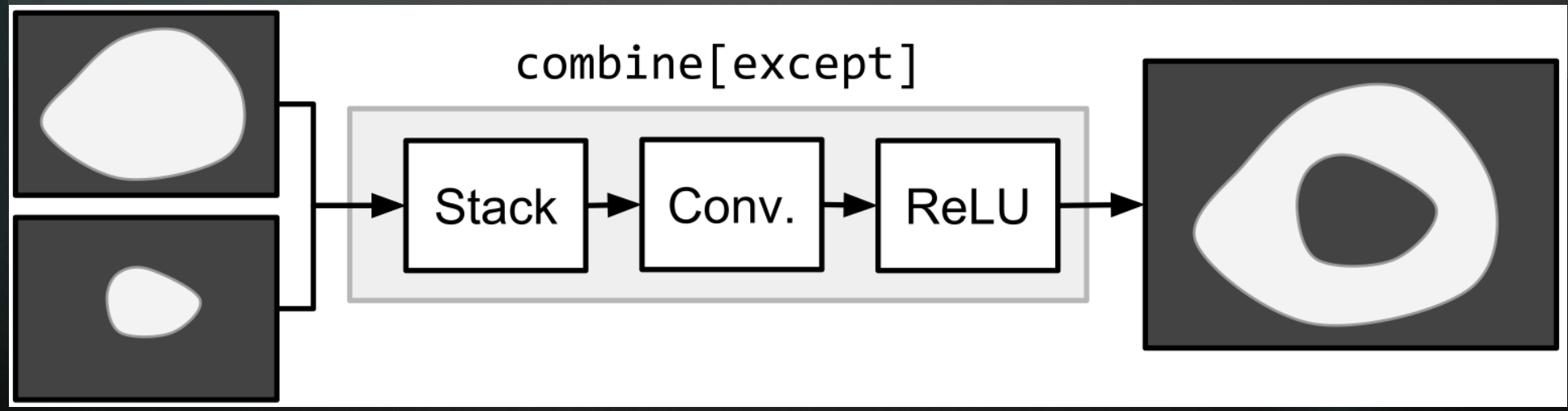
2. The **re-attend** Module for Relationships



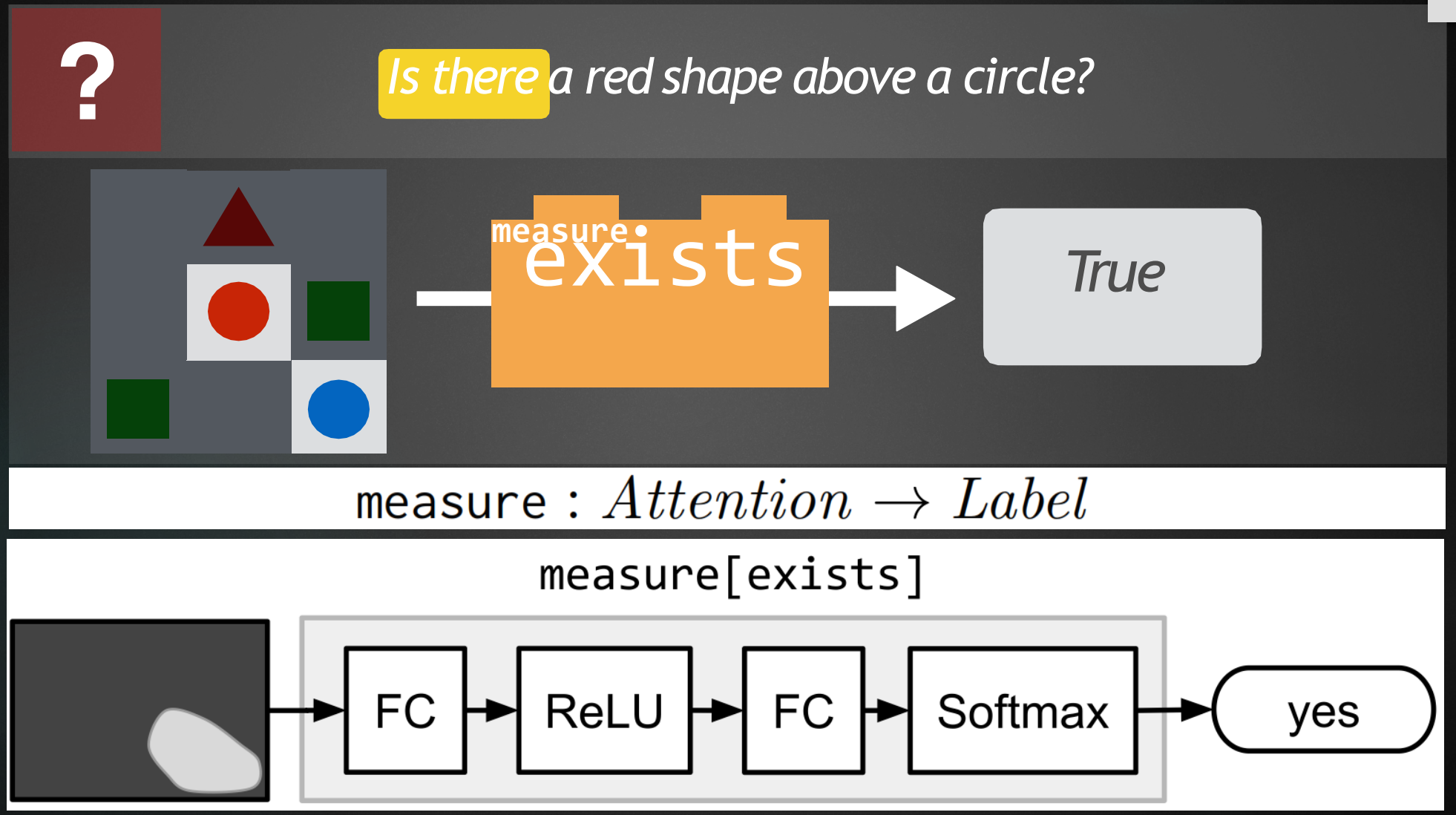
3. The **combine** Module for Logical Operations



$$\texttt{combine} : Attention \times Attention \rightarrow Attention$$



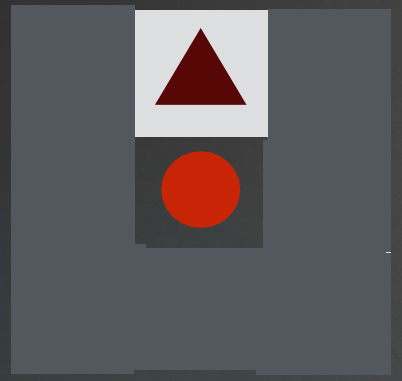
4. The **measure** Module

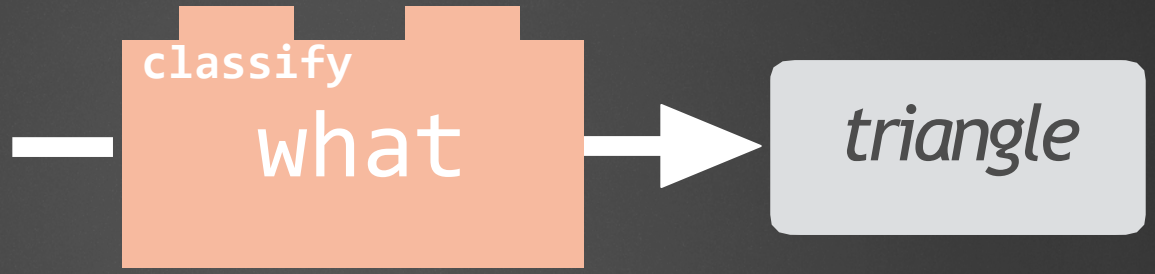


5. The **classify** Module

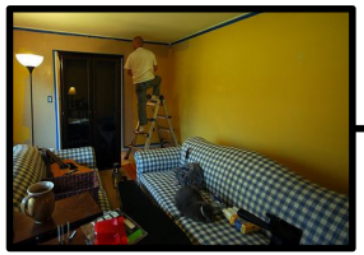
?

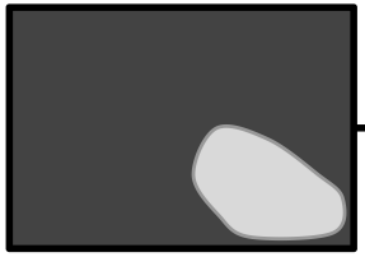
What is the a red shape above a circle?

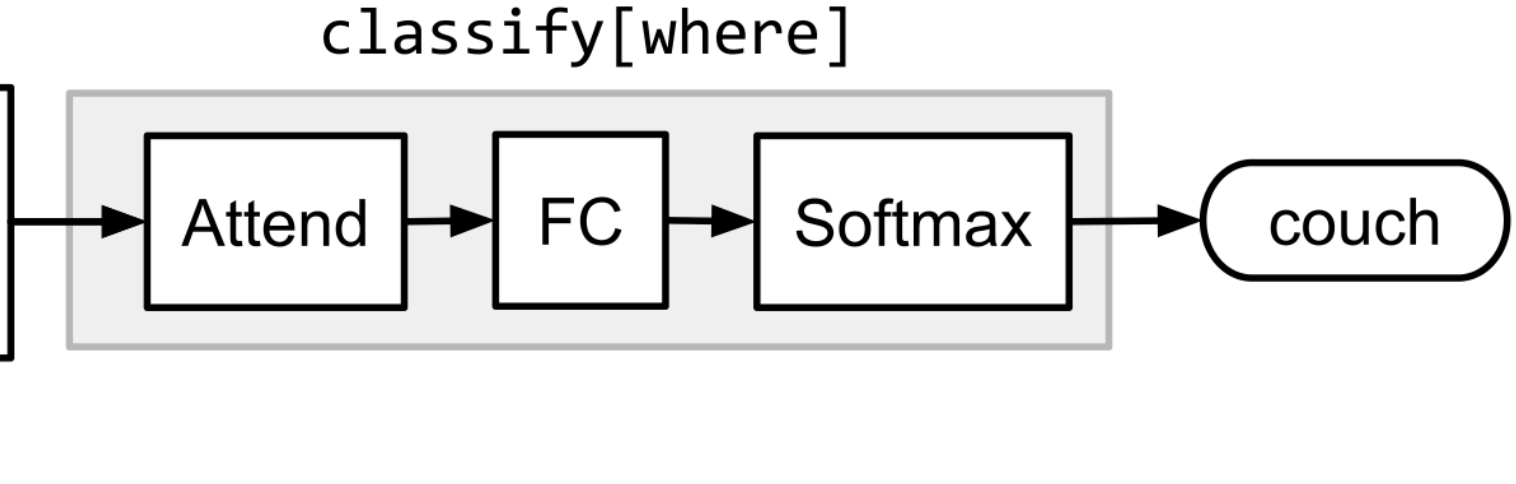




$$\texttt{classify} : Image \times Attention \rightarrow Label$$

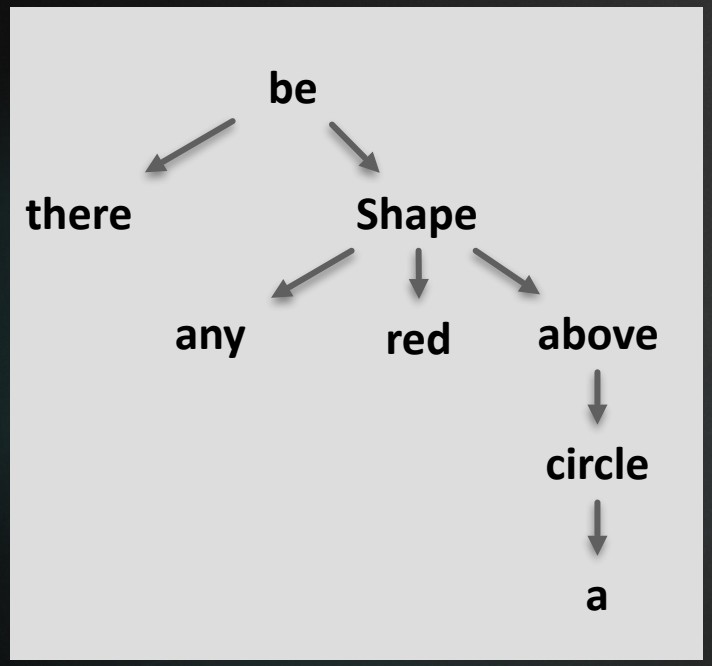


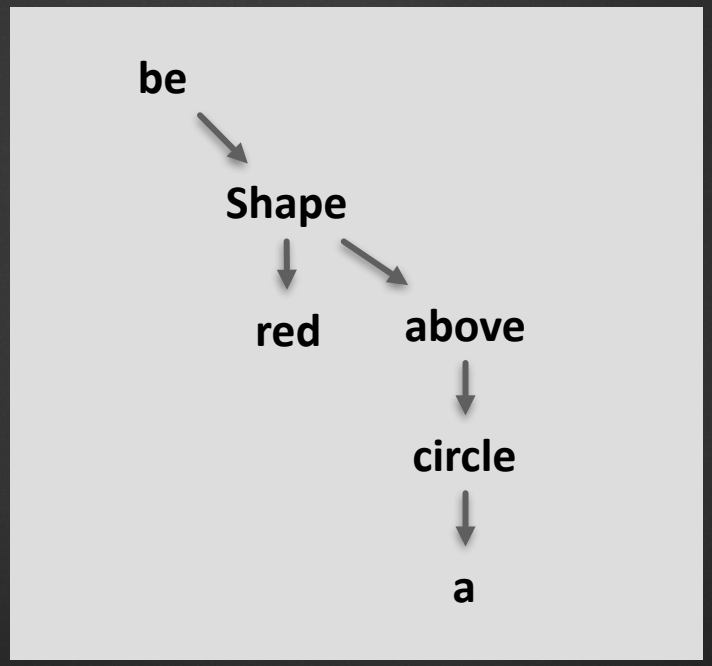


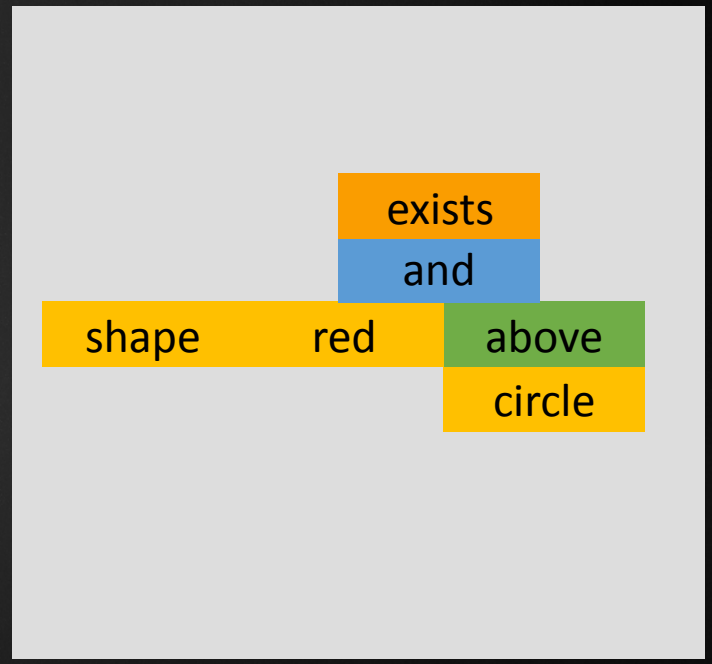


Training NMNs

Is there a red shape above a circle?

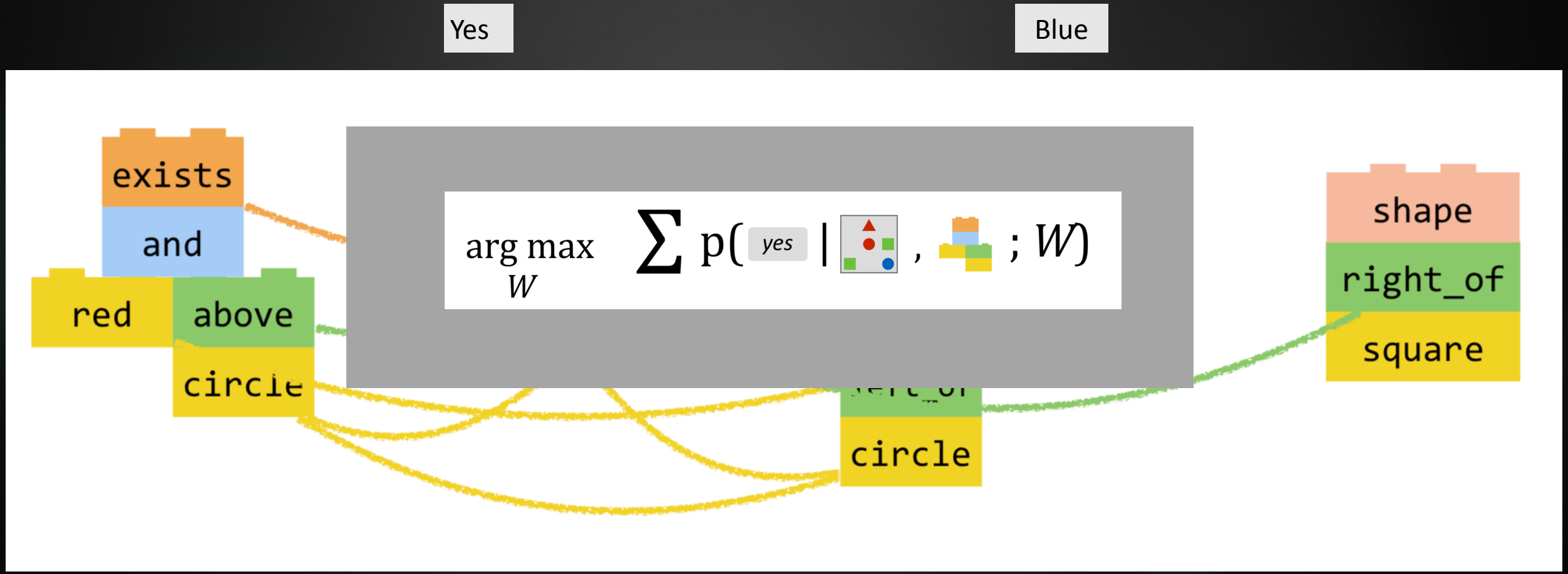






Training NMNs

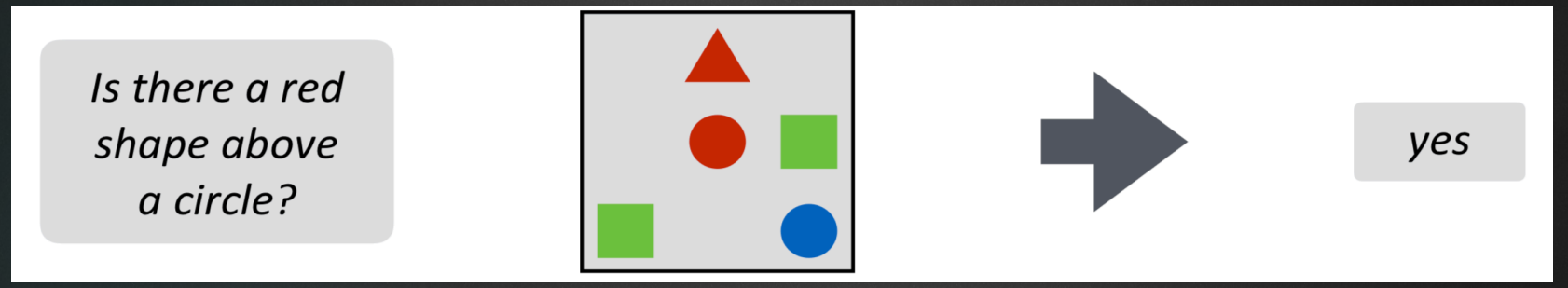
- Learning parameters for individual module
- Supervised learning problem to maximize probability of output

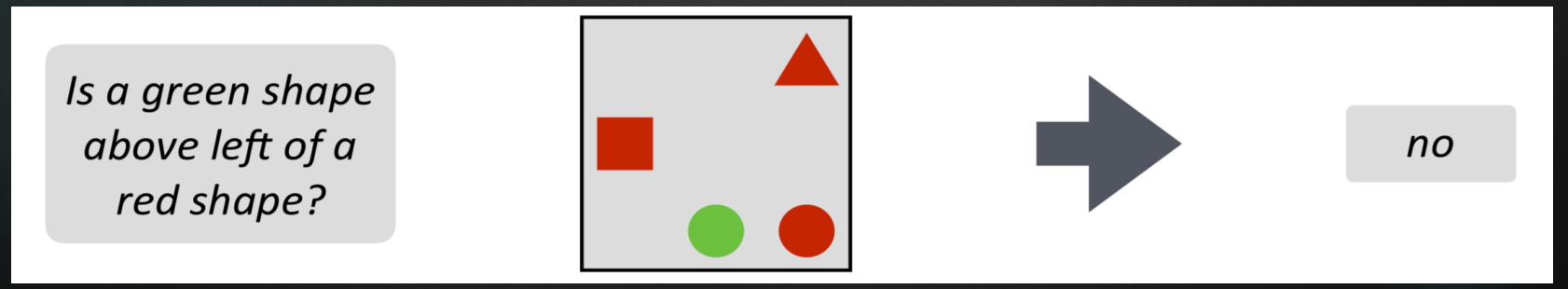


VQA Dataset

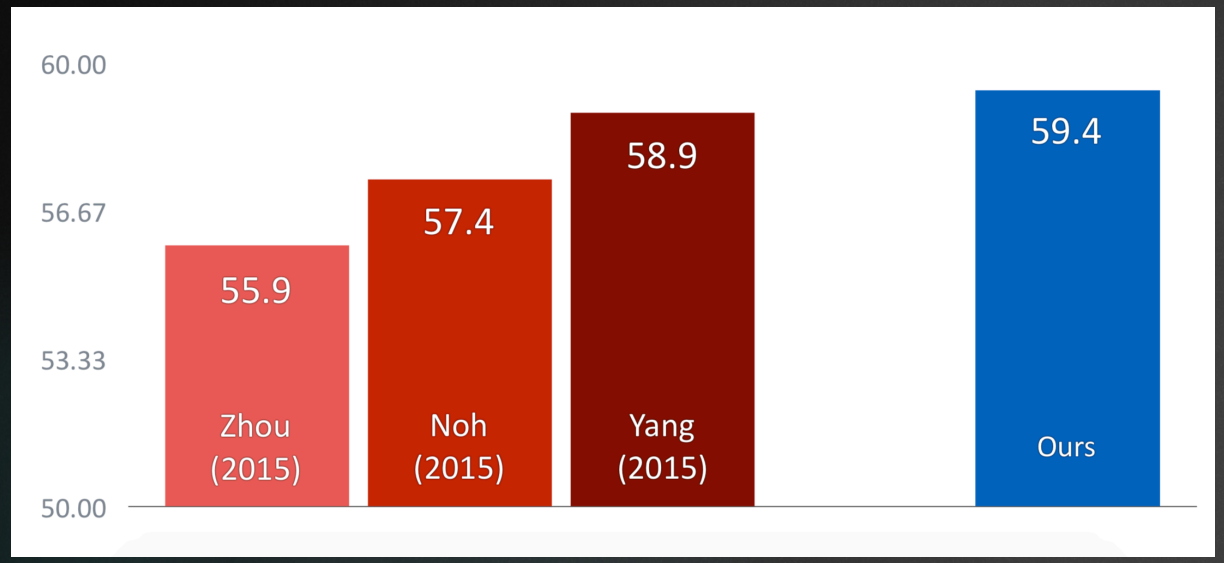
[Antol et al. 2015]

SHAPES Dataset - Novel





Experiment Results - VQA



	test-dev				test
	Yes/No	Number	Other	All	All
LSTM [2]	78.20	35.7	26.6	48.8	–
VIS+LSTM [2]	78.9	35.2	36.4	53.7	54.1
NMN	69.38	30.7	22.7	42.7	–
NMN+LSTM	77.7	37.2	39.3	54.8	**55.1**

What color is she wearing?

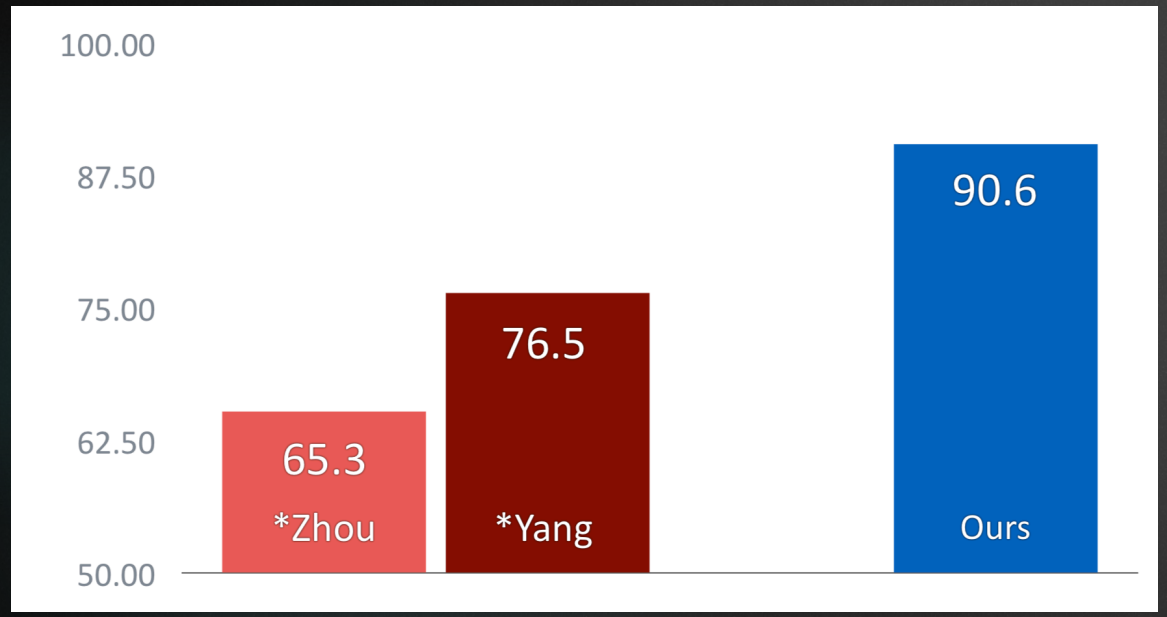
color

wear

white

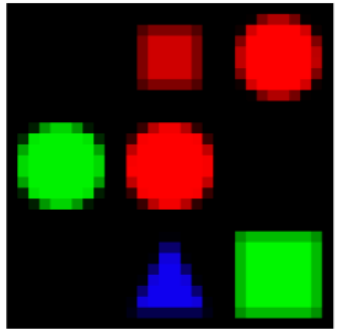
Experiment Results - SHAPES

- Size: number of modules needed to instantiate NMN
- NMN (easy): modified training set with no size-6 questions



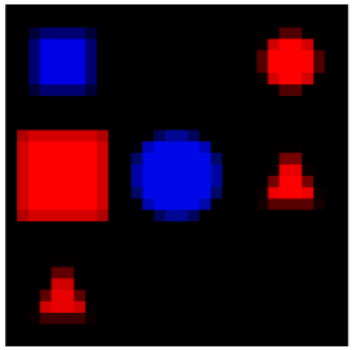
	size 4	size 5	size 6	All
Majority	64.4	62.5	61.7	63.0
VIS+LSTM	71.9	62.5	61.7	65.3
NMN	89.7	92.4	85.2	**90.6**
NMN (easy)	97.7	91.1	89.7	**90.8**

Discussion

		
what color is the vase?	*is the bus full of passengers?*	*is there a red shape above a circle?*

`classify[color](`
`  attend[vase])`
	`measure[is](`
	`  combine[and](`
	`  attend[bus],`
	`  attend[full])`
		`measure[is](`
		`  combine[and](`
		`  attend[red],`
		`  re-attend[above](`
		`  attend[circle])))`

green (green)	yes (yes)	no (no)

		
what material are the boxes made of?	*is this a clock?*	*is a red shape blue?*

`classify[material](`
`  attend[box])`
	`measure[is](`
	`attend[clock])`
		`measure[is](`
		`  combine[and](`
		`  attend[red],`
		`  attend[blue]))`

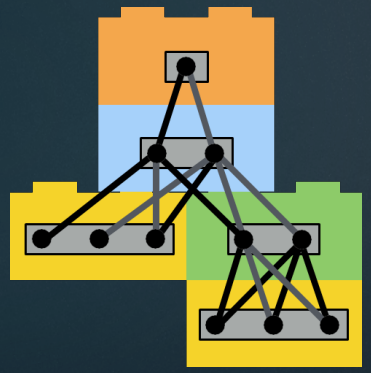
leather (cardboard)	yes (no)	yes (no)


Show apps

A Closer Look…

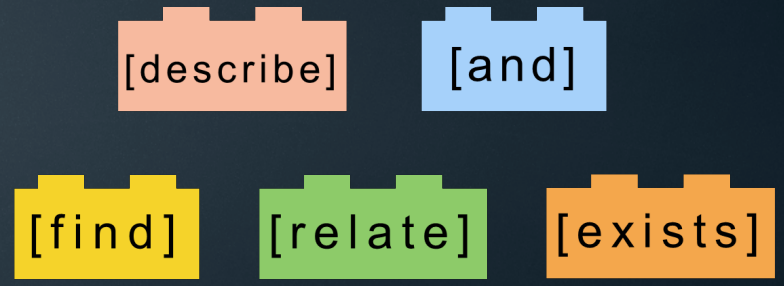
Pros:

- Combines advantages of:
  - Representation learning (like a neural net)
  - Compositionality (like a semantic parser)



Cons:

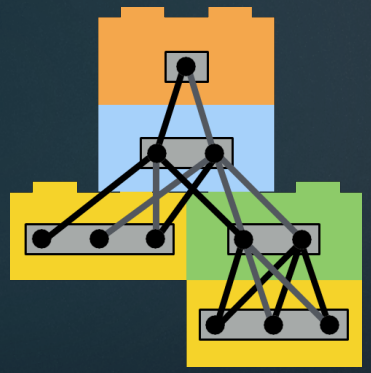
- Limited to produce arbitrary forms



A Closer Look…

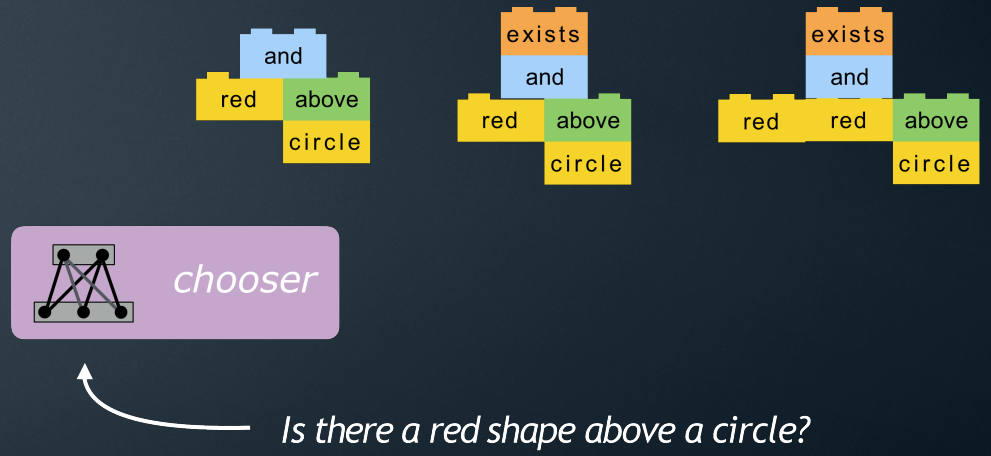
Pros:

- Combines advantages of:
  - Representation learning (like a neural net)
  - Compositionality (like a semantic parser)



Cons:

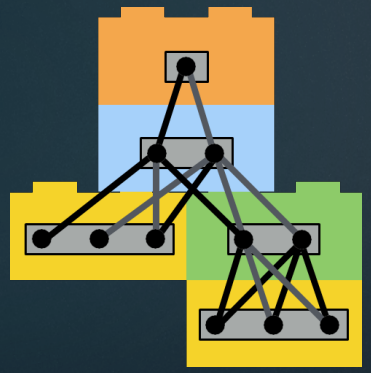
- Limited to produce arbitrary forms
- Structure selection



A Closer Look…

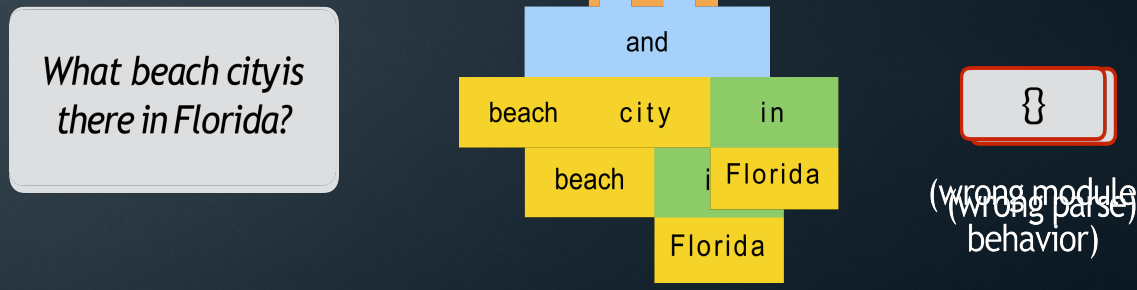
Pros:

- Combines advantages of:
  - Representation learning (like a neural net)
  - Compositionality (like a semantic parser)



Cons:

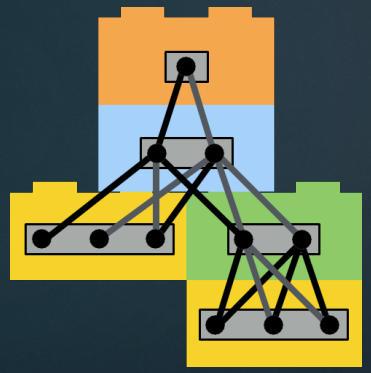
- Limited to produce arbitrary forms
- Structure selection
- Potential error in generated model



A Closer Look…

Pros:

- Combines advantages of:
  - Representation learning (like a neural net)
  - Compositionality (like a semantic parser)

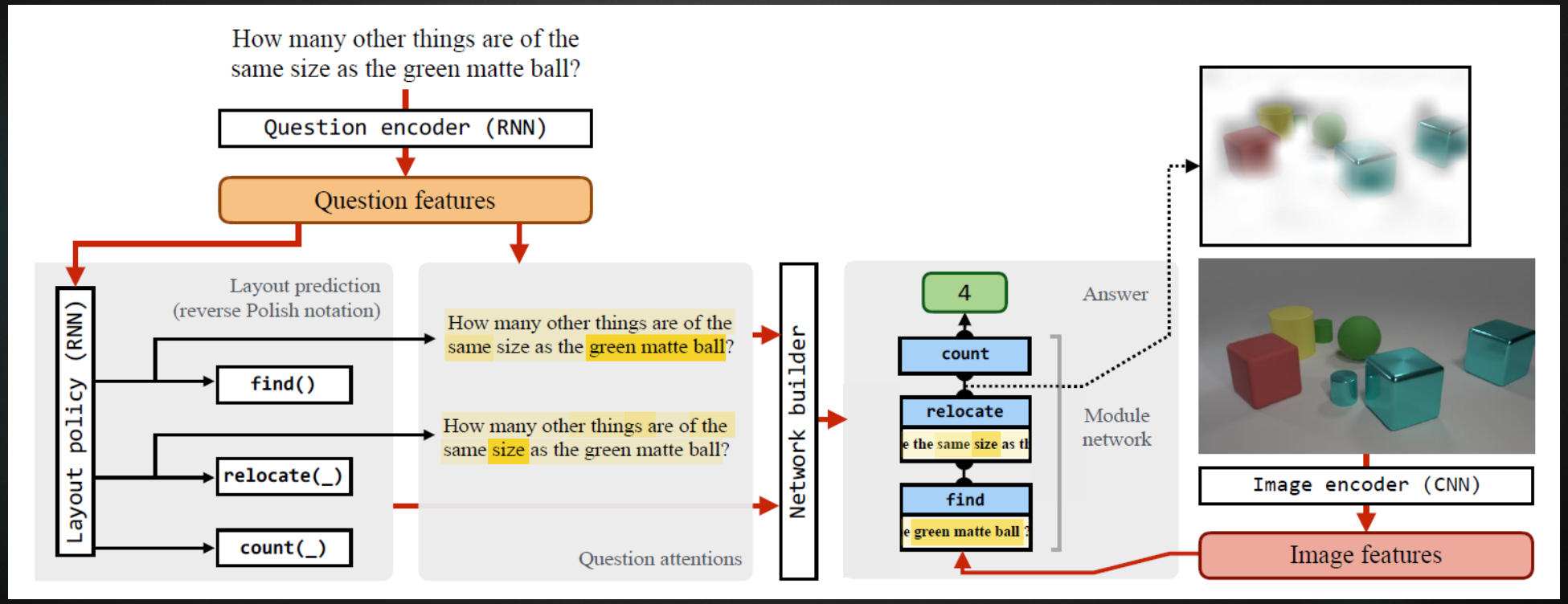


Cons:

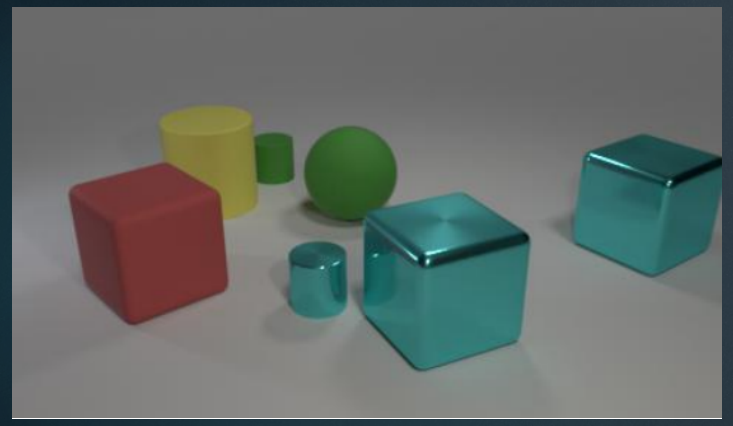
- Limited to produce arbitrary forms
- Structure selection
- Potential error in generated model
- End-2-End structure learning

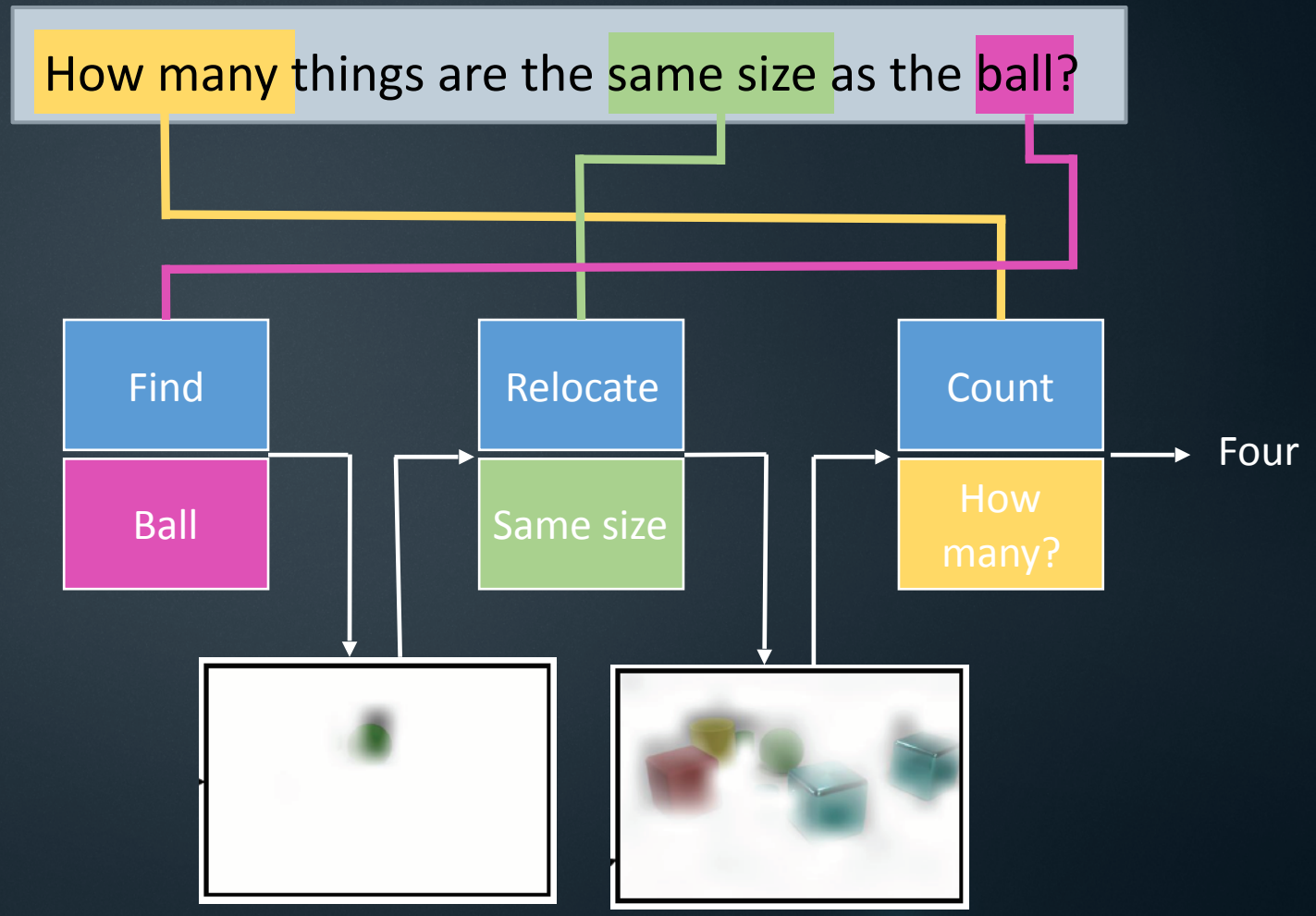
Extension of the Work

- Learning to Reason: End-to-End Module Networks for Visual Question Answering

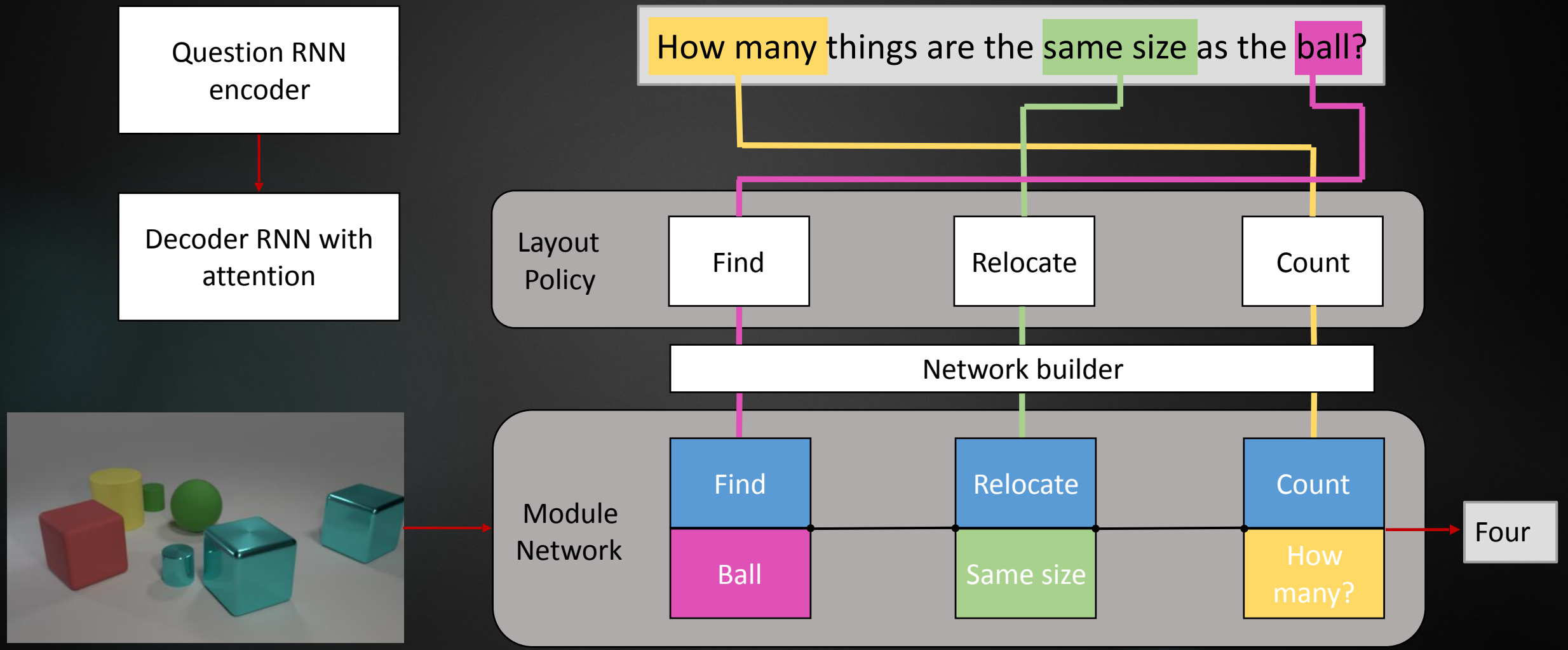


Extension of the Work



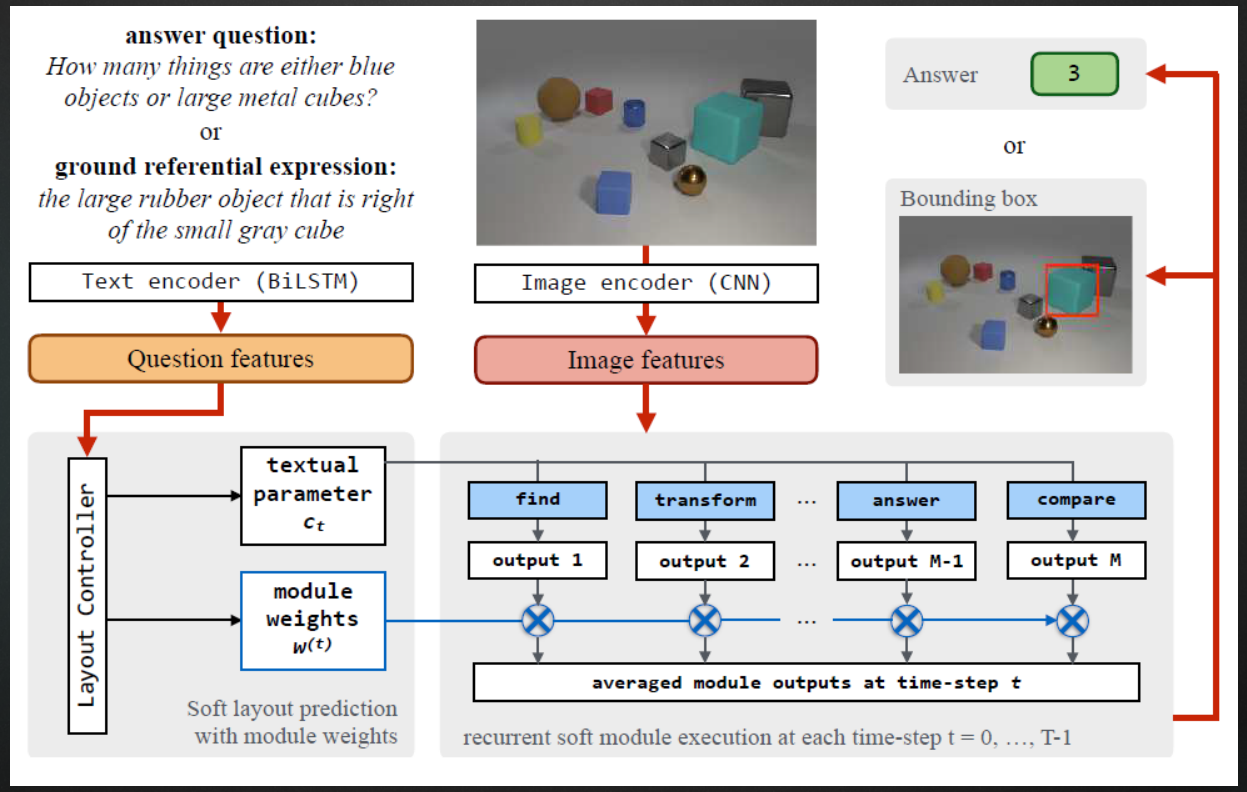


Extension of the Work



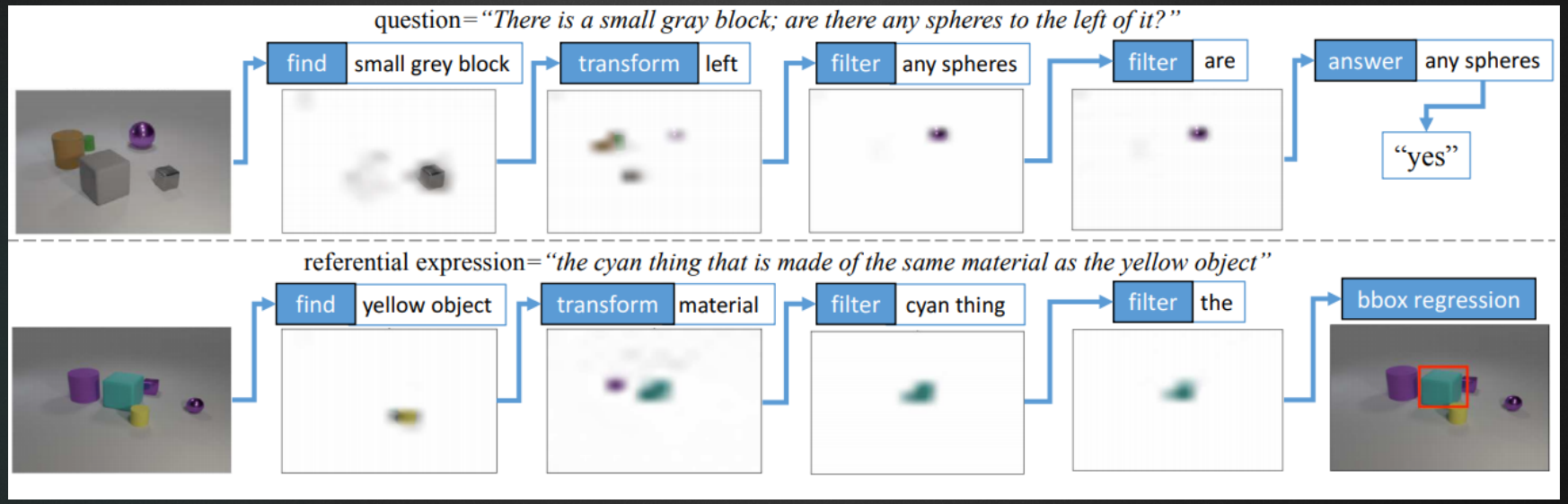
Extension of the Work

- Explainable Neural Computation via Stack Neural Module Networks



Extension of the Work

- Answer visual questions
- Ground referential expressions



Bibliography

- Andreas J, Rohrbach M, Darrell T, Klein D. Neural module networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016 (pp. 39-48).
- Andreas J, Rohrbach M, Darrell T, Klein D. Learning to compose neural networks for question answering. 2016. PowerPoint Presentation.
- Chen L, Tang S. Visual Question Answering. 2016. PowerPoint Presentation.
- Wu Q, Teney D, Wang P, Shen C, Dick A, van den Hengel A. Visual question answering: A survey of methods and datasets. Computer Vision and Image Understanding. 2017 Oct 1;163:21-40.
- Hu R, Andreas J, Darrell T, Saenko K. Explainable neural computation via stack neural module networks. In Proceedings of the European Conference on Computer Vision (ECCV) 2018 (pp. 53-69).

Thank You

and Happy Spring