Neural Modular Networks

Joseph E. Gonzalez
Co-director of the RISE Lab
jegonzal@cs.berkeley.edu

Today

Deep Compositional Question Answering with Neural Module Networks

Jacob Andreas Marcus Rohrbach Trevor Darrell Dan Klein
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
{jda,rohrbach,trevor,klein}@{cs,eecs,eecs,cs}.berkeley.edu

Abstract

Visual question answering is fundamentally compositional in nature—a question like where is the dog? *shares substructure with questions like* what color is the dog? *and* where is the cat? *This paper seeks to simultaneously exploit the representational capacity of deep networks and the compositional linguistic structure of questions. We describe a procedure for constructing and learning* neural module networks*, which compose collections of jointly-trained neural "modules" into deep networks for question answering. Our approach decomposes questions into their linguistic substructures, and uses these structures to dynamically instantiate modular networks (with reusable components for recognizing dogs, classifying colors, etc.). The resulting compound networks are jointly trained. We evaluate our approach on two challenging datasets for visual question answering, achieving state-of-the-art results on both the VQA natural image dataset and a new dataset of complex questions about abstract shapes.*

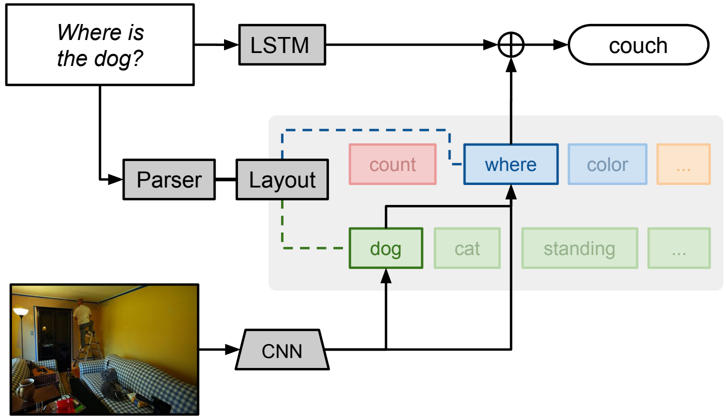


Figure 1: A schematic representation of our proposed model—the shaded gray area is a *neural module network* of the kind introduced in this paper. Our approach uses a natural language parser to dynamically lay out a deep network composed of reusable modules. For visual question answering tasks, an additional sequence model provides sentence context and learns common-sense knowledge.

What Problem is being solved?

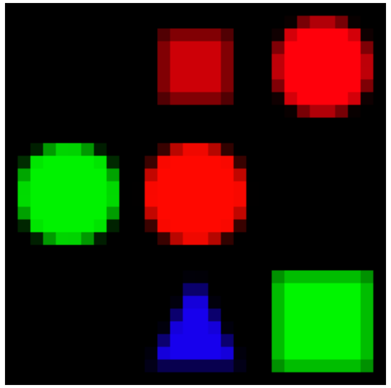
- **Problem Domain:** Visual Question Answering
  - “Visual Turing test”

how many different lights in various different shapes and sizes?

what is the color of the horse?

what color is the vase?

is the bus full of passengers?

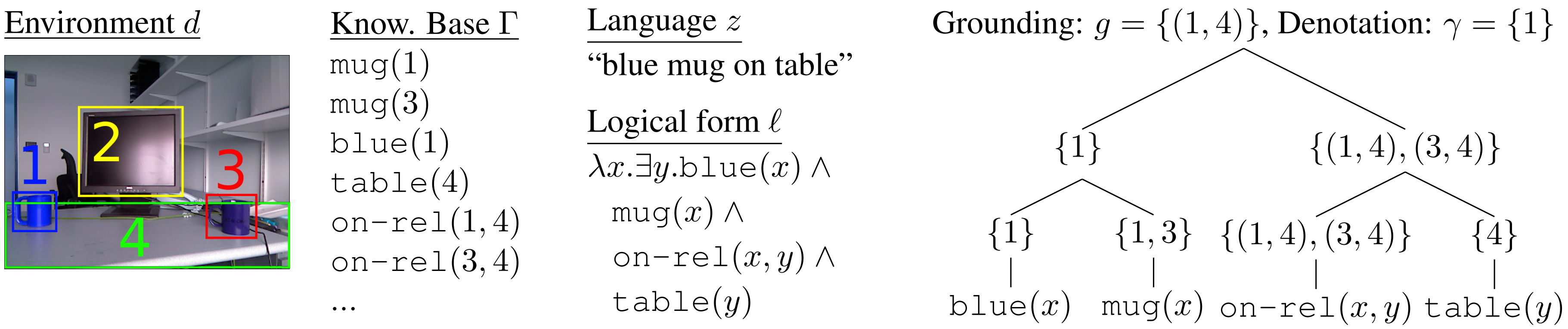


is there a red shape above a circle?

Prior State of the Art

➢ Semantic parsing and logic:

Jointly Learning to Parse and Perceive: Connecting Natural Language to the Physical World



(a) Perception f_{per} produces a logical knowledge base Γ from the environment d using an independent classifier for each category and relation.

(b) Semantic parsing f_{prs} maps language z to a logical form ℓ.

(c) Evaluation f_{eval} evaluates a logical form ℓ on a logical knowledge base Γ to produce a grounding g and denotation γ.

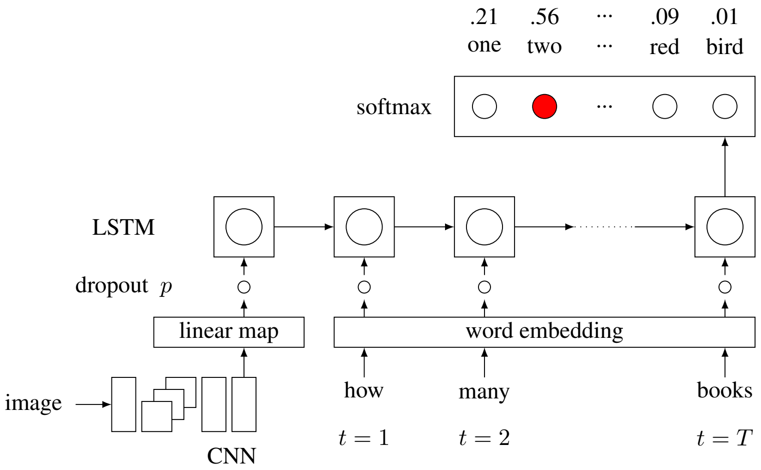
Figure 2: Overview of Logical Semantics with Perception (LSP).

➢ Dependent on pre-trained computer vision models to populate database

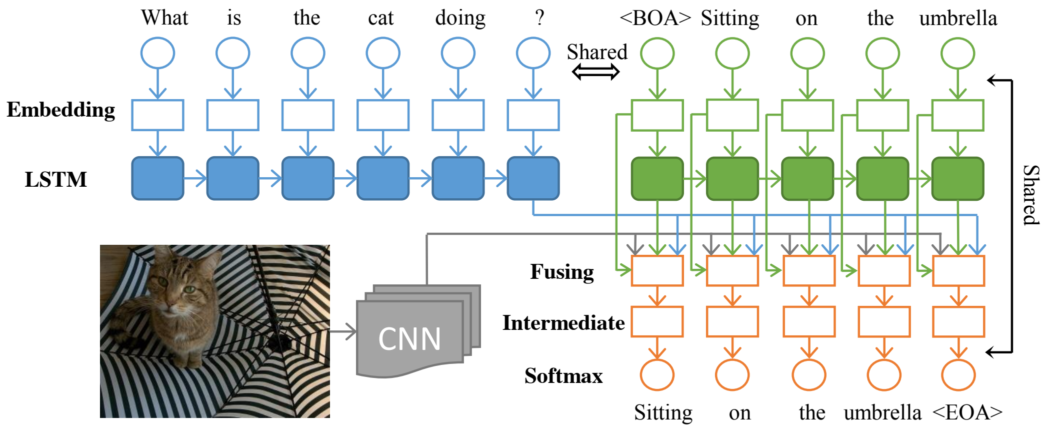
Prior State of the Art

- Deep Embeddings

Image Question Answering: A Visual Semantic Embedding Model and a new Dataset

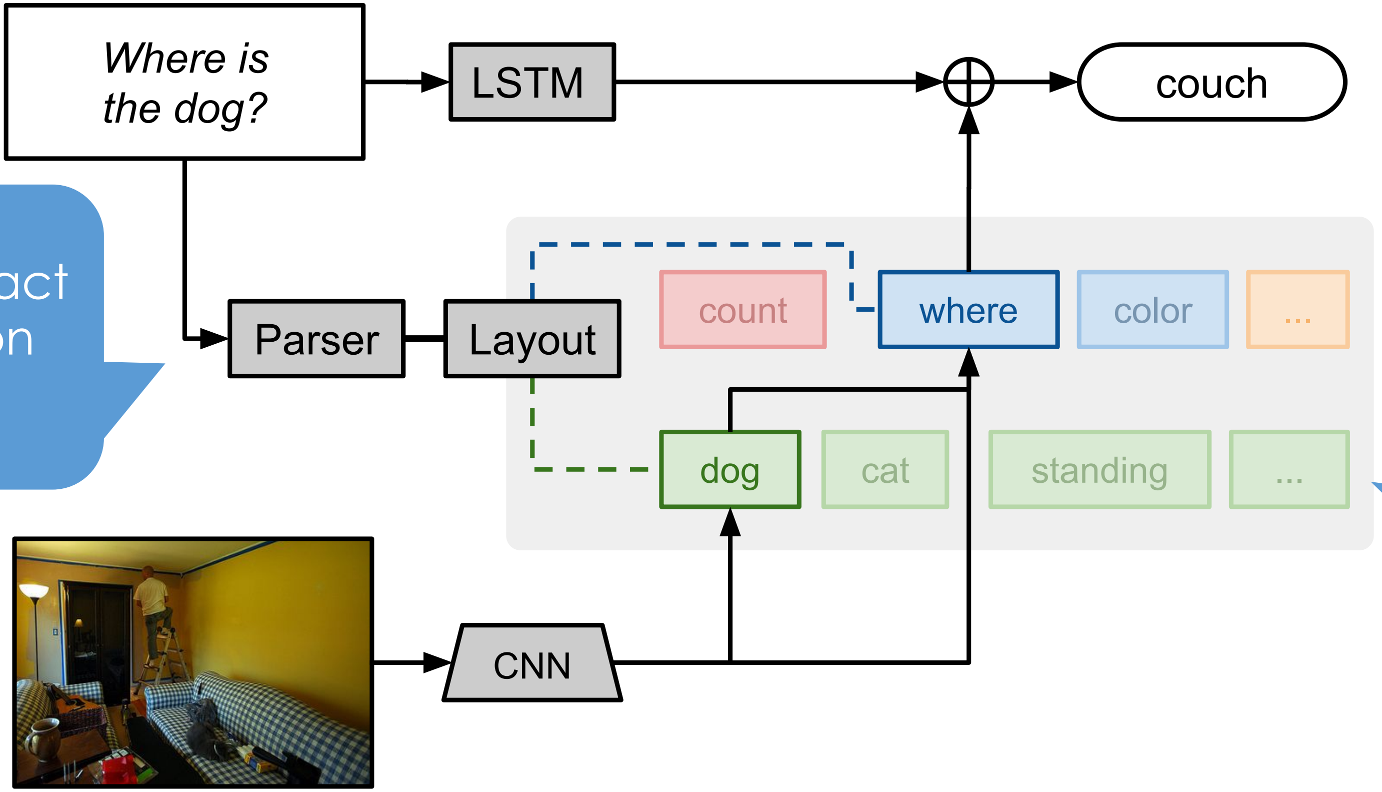


Are you Talking to a Machine? Datasets and Methods for Multilingual Image Question Answering



- Learned end-to-end but image representation is independent of the question.

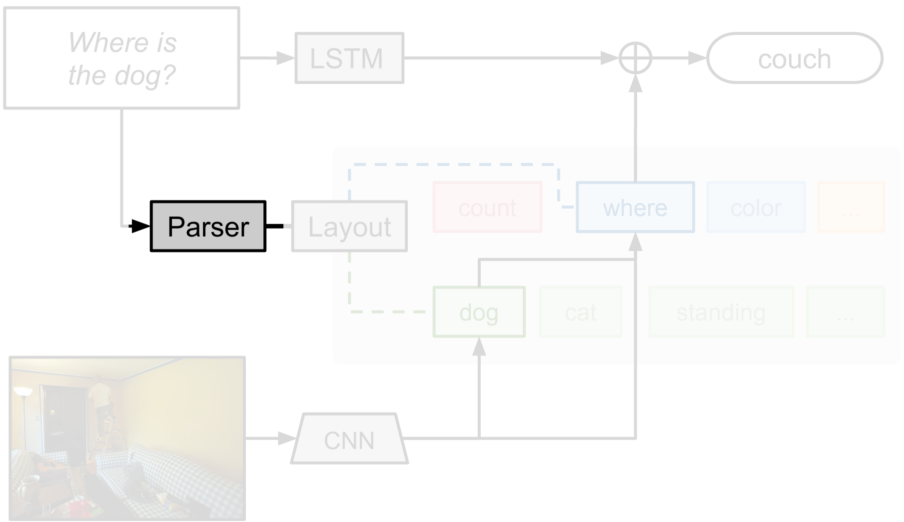
Proposed Solution



Use parser to extract logical expression from question.

Compose a neural network to answer question

Reuse network components across problems



Question:

Is there a circle next to a square?

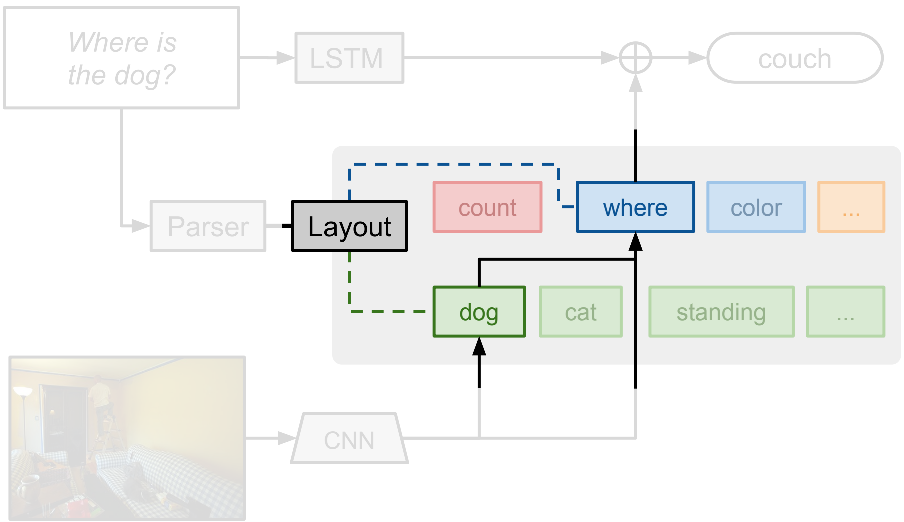
Logical Expression:

`**is**(circle, **next-to**(square))`

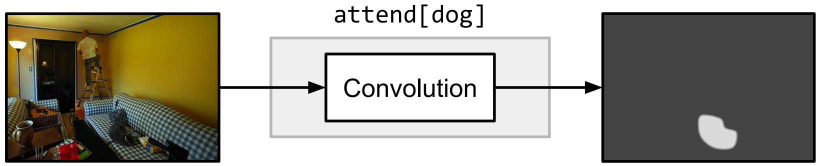
- **Objective:** Convert question into logical expression.
- Conceptually → Inducing a ***program*** from a question
- Also probably the more brittle part of the work
  - Addressed in follow-up paper
  - Alternative solution: user writes logical expression → programming

Neural Modules

"Learned Sub-routines/Functions"

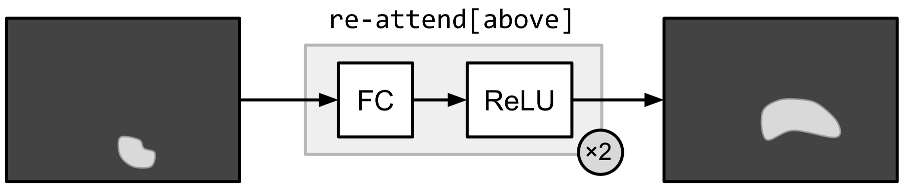


$\texttt{attend} : \mathit{Image} \rightarrow \mathit{Attention}$

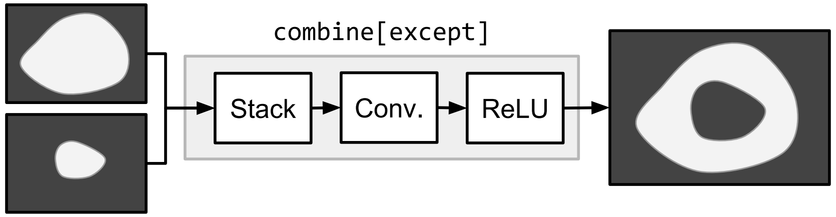


Separate weights for each argument e.g., [dog]

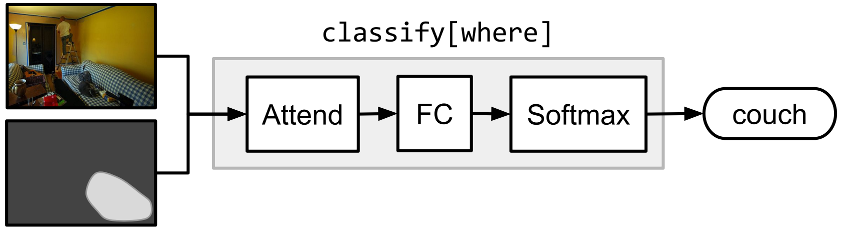
$\texttt{re-attend} : \mathit{Attention} \rightarrow \mathit{Attention}$



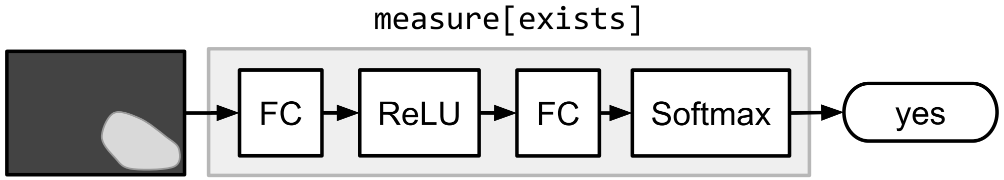
$\texttt{combine} : \mathit{Attention} \times \mathit{Attention} \rightarrow \mathit{Attention}$

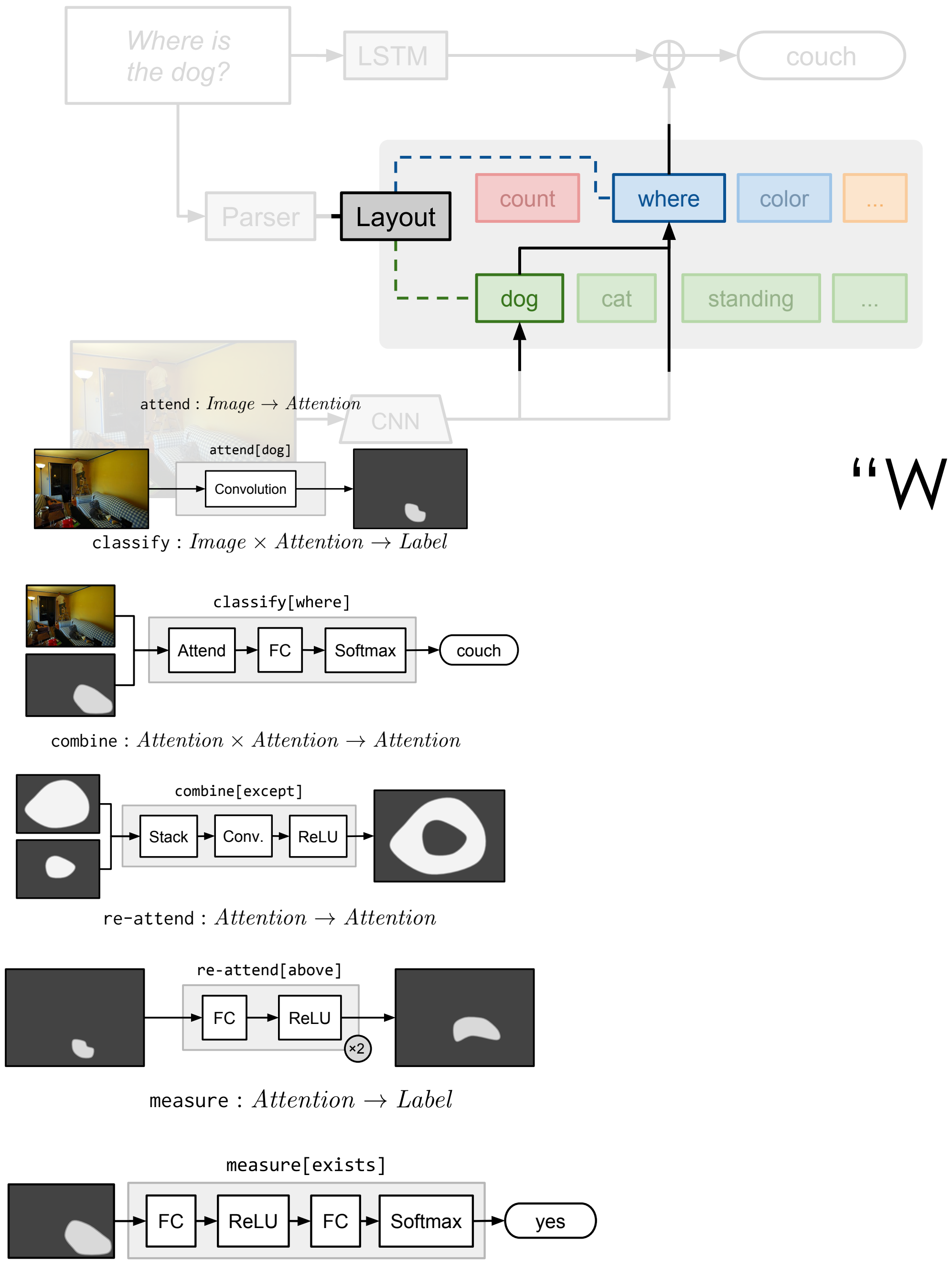


$\texttt{classify} : \mathit{Image} \times \mathit{Attention} \rightarrow \mathit{Label}$



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

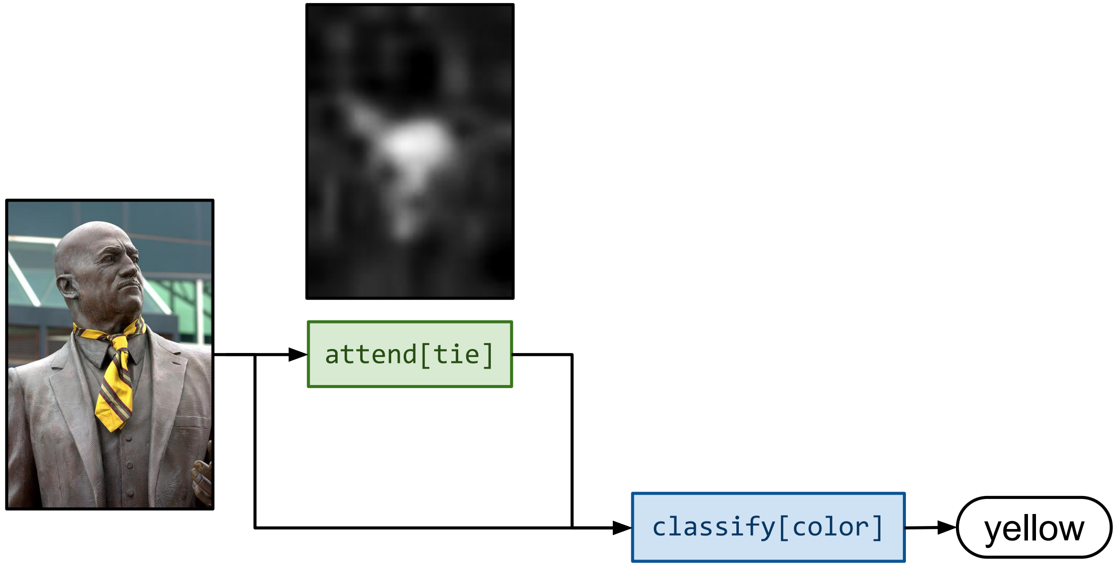


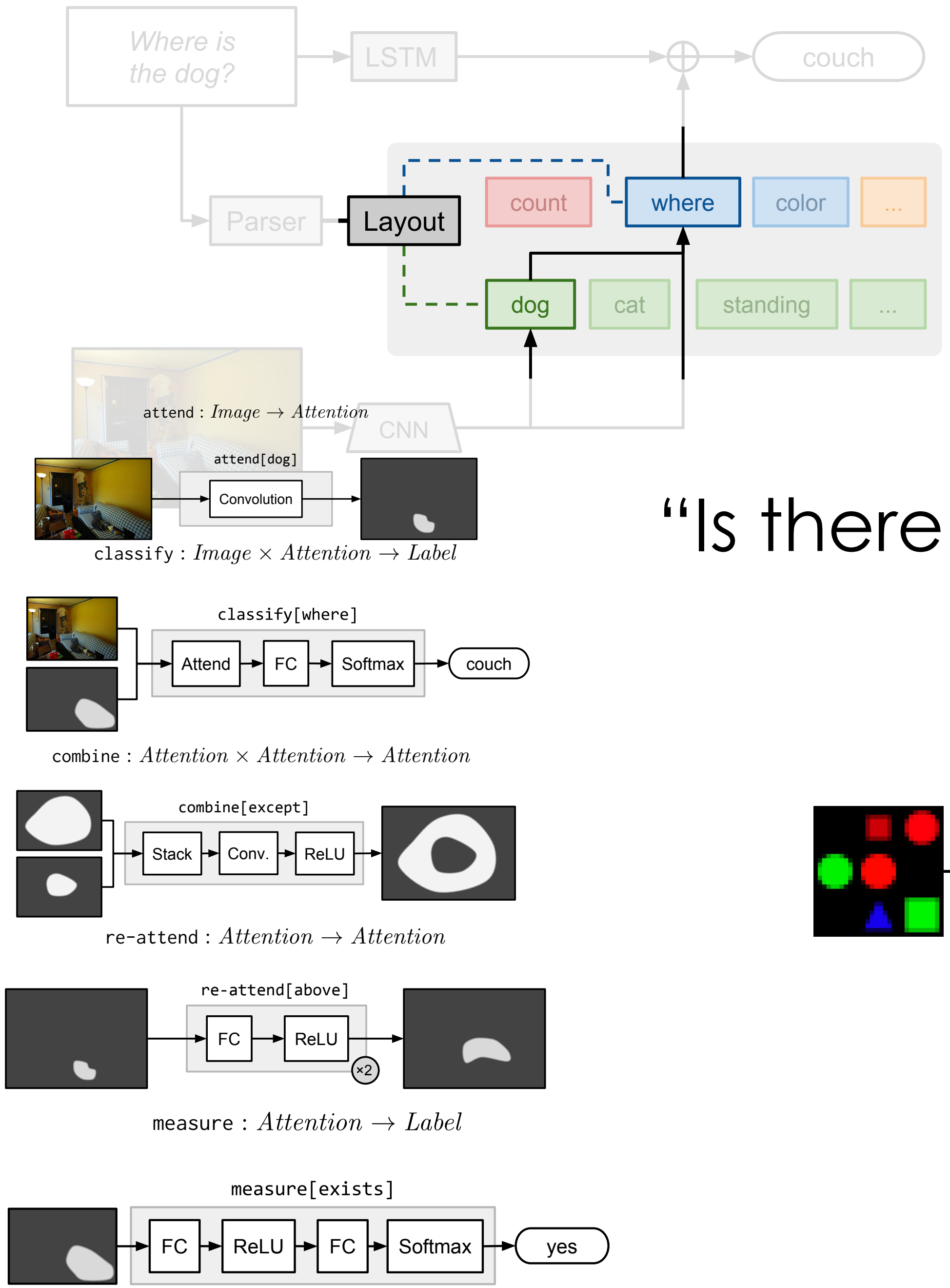


Composition!

"Learned programs"

"What color is his tie?"

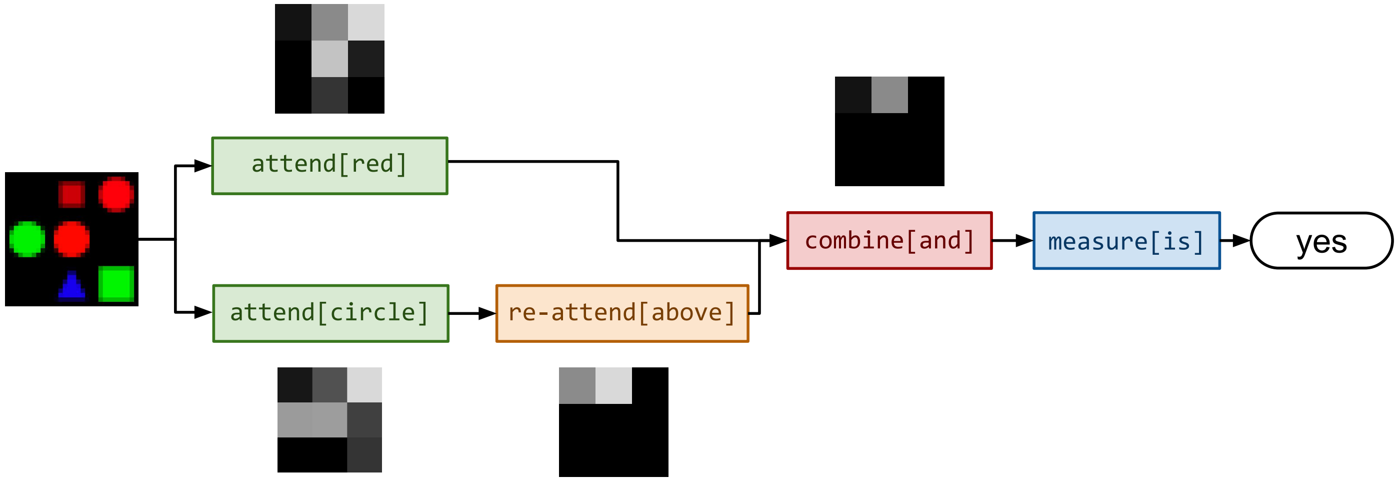




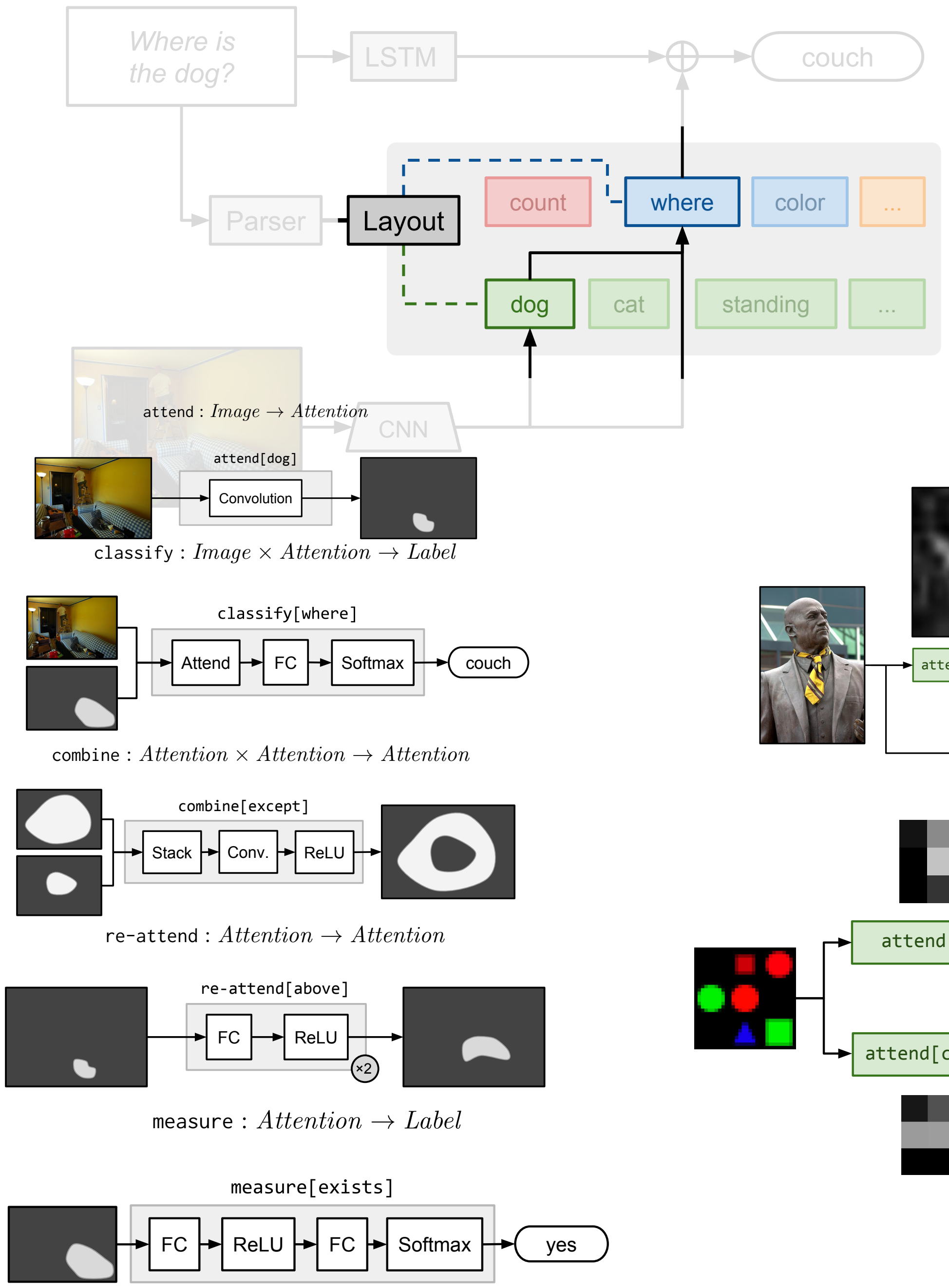
Composition!

"Learned programs"

"Is there a red shape above a circle?"

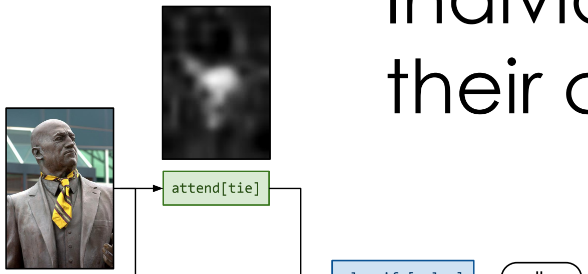


Training

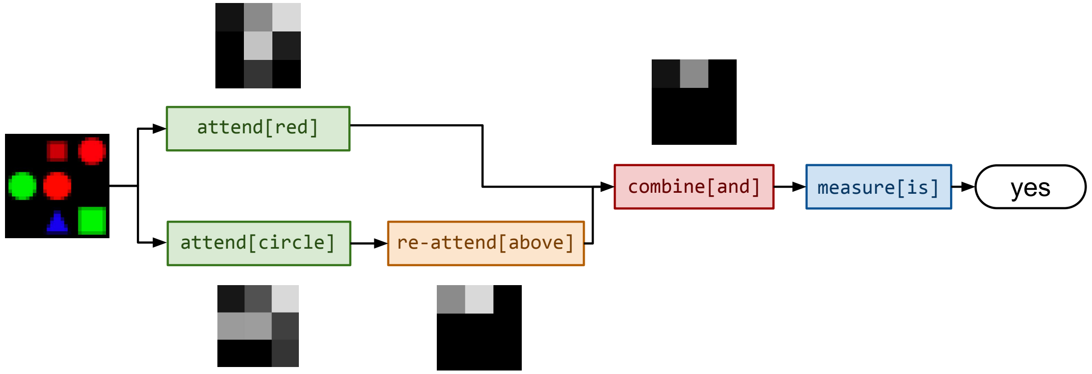


Train multiple graphs at once with shared modules.

Individual models learn through their composition.



No pre-training



Evaluation Metrics and Results

- Accuracy on VQA benchmarks
  - Existing benchmarks only require limited reasoning…
- Introduce new Shapes Benchmark

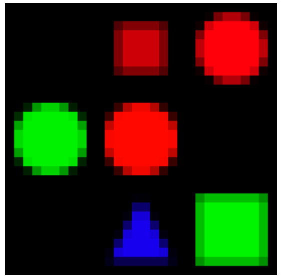
Shapes Benchmark

	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**

VQA Benchmark

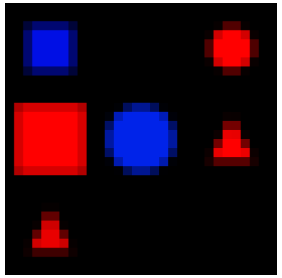
	test-dev				test
	Yes/No	Number	Other	All	All
LSTM [2]	78.20	35.7	26.6	48.8	–
VIS+LSTM [2]	78.9	35.2	36.4	53.7	54.1
NMN	69.38	30.7	22.7	42.7	–
NMN+LSTM	77.7	37.2	39.3	54.8	**55.1**

Qualitative Results

				
how many different lights in various different shapes and sizes?	*what is the color of the horse?*	*what color is the vase?*	*is the bus full of passengers?*	*is there a red shape above a circle?*

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

four (four)	brown (brown)	green (green)	yes (yes)	no (no)

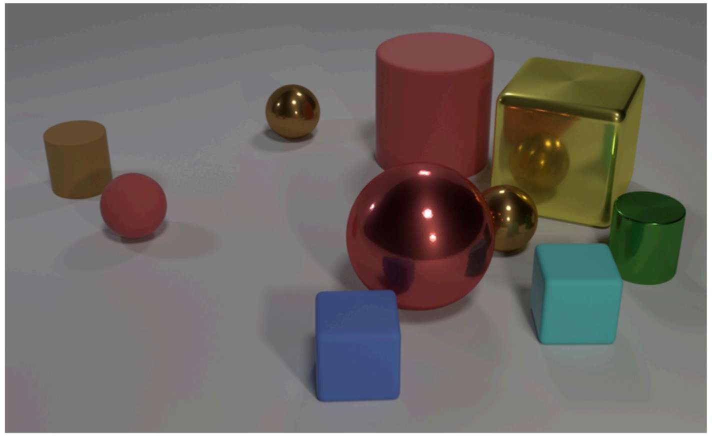
				
what is stuffed with toothbrushes wrapped in plastic?	*where does the tabby cat watch a horse eating hay?*	*what material are the boxes made of?*	*is this a clock?*	*is a red shape blue?*

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

container (cup)	pen (barn)	leather (cardboard)	yes (no)	yes (no)

Impact

- Over 300 citations (pretty good)
- Follow-up work "*Learning to Reason: End-to-End Module Networks for Visual Question Answering*" address limitations of parsing.
  - Uses Policy RNN to predict composition (trained using RL)



Q: Are there an **equal number** of **large things** and **metal spheres**?
Q: **What size** is the **cylinder that is left of** the **brown metal** thing **that is left of** the **big sphere**?
Q: There is a **sphere** with the **same size as** the **metal cube**; is it **made of the same material as** the **small red sphere**?

Method	**Overall**	Exist	Count	Compare Integer			Query Attrib		
				equal	less	more	size	color	mater
CNN+BoW [26]	48.4	59.5	38.9	50	54	49	56	32	58
CNN+LSTM [4]	52.3	65.2	43.7	57	72	69	59	32	58
CNN+LSTM+MCB [9]	51.4	63.4	42.1	57	71	68	59	32	57
CNN+LSTM+SA [25]	68.5	71.1	52.2	60	82	74	87	81	88
NMN (expert layout) [3]	72.1	79.3	52.5	61.2	77.9	75.2	84.2	68.9	82.
ours - policy search from scratch	69.0	72.7	55.1	71.6	85.1	79.0	88.1	74.0	86.
ours - cloning expert	78.9	83.3	63.3	68.2	87.2	85.4	90.5	80.2	88.
ours - policy search after cloning	**83.7**	**85.7**	**68.5**	**73.8**	**89.7**	**87.7**	**93.1**	**84.8**	**91.**

Points to a bigger opportunity…

- **Composition of learned modules**
- **Conjecture:** Increasing "non-experts" will compose existing ML models to solve new complex problems.
  - Organizations will develop and **reuse** model components in multiple tasks
  - Training will span many different **neural module programs**
- **Needed?**
  - **Abstractions** for individual components
  - **Mechanisms** for composition and joint training