Visual Question Answering

Soham Parikh

Computer and Information Science
University of Pennsylvania

4/18/19

Motivation for Attention

Previous Approaches: Use a summary of the context (image/passage)

However, some parts of the context are more important to answer the question!

Context

The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the "Top Gear" host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broadcaster found he had subjected producer Oisin Tymon "to an unprovoked physical and verbal attack." ...

Query

Producer **X** will not press charges against Jeremy Clarkson, his lawyer says.

Answer

Oisin Tymon

Q: what is the color of the bird? **A**: **white**

CNN/Daily Mail dataset

Teaching Machines to Read and Comprehend (NIPS 2015)

Context :

> @entity4 if you feel a ripple in the force today , it may be the news that the official @entity6 is getting its first gay character . according to the sci-fi website @entity9 , the upcoming novel "@entity11" will feature a capable but flawed @entity13 official named @entity14 who "also happens to be a lesbian." the character is the first gay figure in the official @entity6 – the movies , television shows , comics and books approved by @entity6 franchise owner @entity22

Articles collected from CNN and Daily Mail websites.

Replace a named-entity in the article with a placeholder to form the question.

CNN/Daily Mail dataset

Teaching Machines to Read and Comprehend (NIPS 2015)

Context :

@entity4 if you feel a ripple in the force today , it may be the news that the official @entity6 is getting its first gay character . according to the sci-fi website @entity9 , the upcoming novel "@entity11" will feature a capable but flawed @entity13 official named @entity14 who "also happens to be a lesbian." the character is the first gay figure in the official @entity6 – the movies , television shows , comics and books approved by @entity6 franchise owner @entity22

Articles collected from CNN and Daily Mail websites.

Replace a named-entity in the article with a placeholder to form the question.

Query : characters in " @placeholder " movies have gradually become more diverse

CNN/Daily Mail dataset

Teaching Machines to Read and Comprehend (NIPS 2015)

Context :

@entity4 if you feel a ripple in the force today , it may be the news that the official @entity6 is getting its first gay character . according to the sci-fi website @entity9 , the upcoming novel "@entity11" will feature a capable but flawed @entity13 official named @entity14 who "also happens to be a lesbian." the character is the first gay figure in the official @entity6 – the movies , television shows , comics and books approved by @entity6 franchise owner @entity22

Articles collected from CNN and Daily Mail websites.

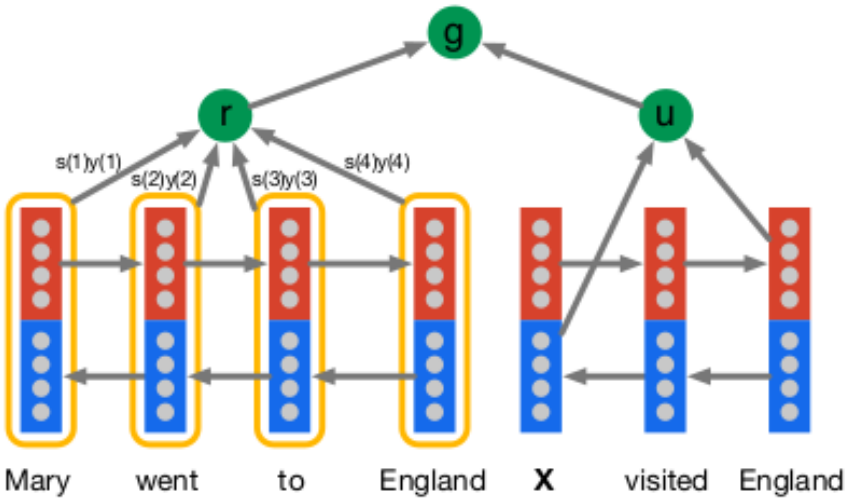
Replace a named-entity in the article with a placeholder to form the question.

Query : characters in " @placeholder " movies have gradually become more diverse

Answer : @entity6

Stanford Attentive Reader

A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task (ACL 2016)



$$y_d(t) = \overrightarrow{y_d}(t) \parallel \overleftarrow{y_d}(t)$$

$$y_q = \overrightarrow{y_q}(|Q|) \parallel \overleftarrow{y_q}(1)$$

$$s(t) = softmax(y_q W_{att} y_d(t))$$

$$g(D, Q) = \sum_t s(t) y_d(t)$$

$$A = \underset{a \in D \cap V}{argmax} W_a g(D, Q)$$

- y_d and y_q obtained from separate BiLSTMs
- Uses bilinear similarity to compute attention $s(t)$
- W_a is a learnable vector

Note: Vocabulary contains tokens for "@entityX"

Stanford Attentive Reader

A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task (ACL 2016)

Model	CNN		Daily Mail	
	Dev	Test	Dev	Test
Frame-semantic model †	36.3	40.2	35.5	35.5
Word distance model †	50.5	50.9	56.4	55.5
Deep LSTM Reader †	55.0	57.0	63.3	62.2
Attentive Reader †	61.6	63.0	70.5	69.0
Impatient Reader †	61.8	63.8	69.0	68.0
MemNNs (window memory) ‡	58.0	60.6	N/A	N/A
MemNNs (window memory + self-sup.) ‡	63.4	66.8	N/A	N/A
MemNNs (ensemble) ‡	66.2*	69.4*	N/A	N/A
Ours: Classifier	67.1	67.9	69.1	68.3
Ours: Neural net	**72.4**	**72.4**	**76.9**	**75.8**

Using Attention for VQA

- Attend over localized regions in an image
- Attend over question words

Q: what is the color of the bird? **A**: **white**

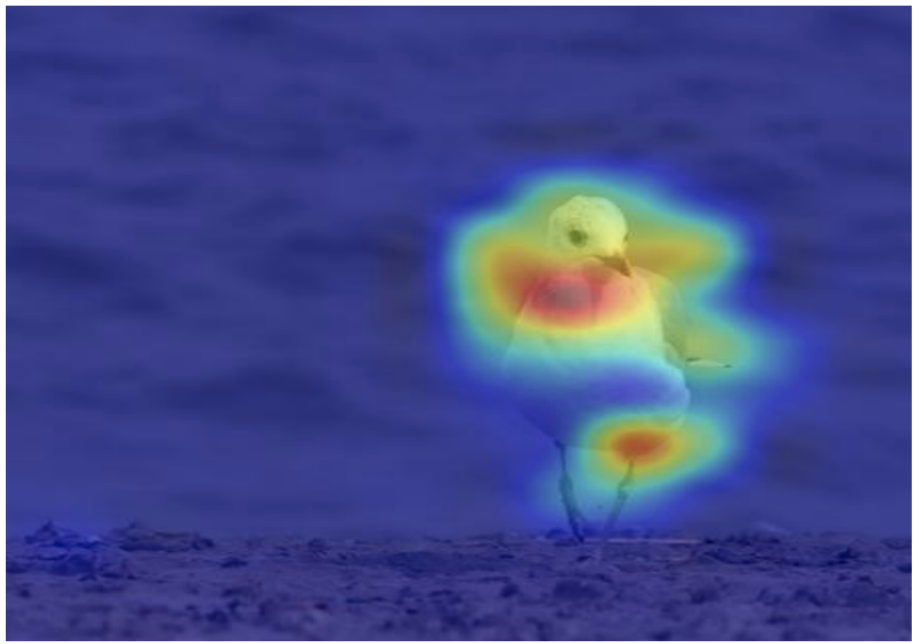


Figure: Example from COCO-QA dataset

Using Attention for VQA

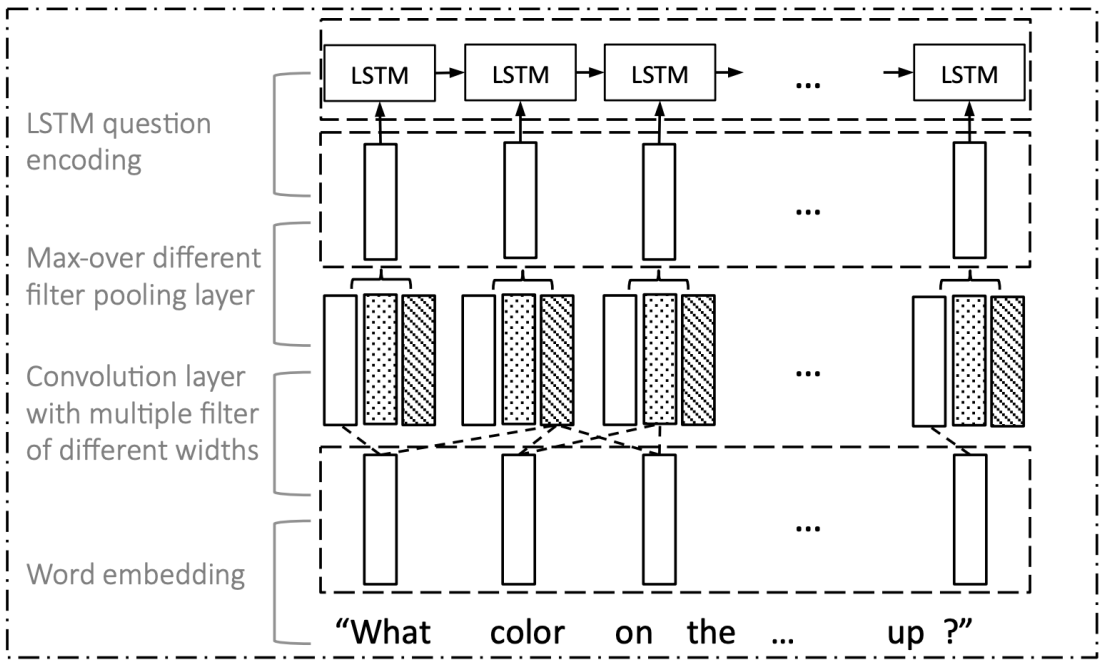
Attending over localized Regions: Features of an image from a CNN layer are of the form $H \times W \times C$, where C is the number of channels.

$\therefore$ There are $H \times W$ location features, each of dimension C. Each of these $N = H \times W$ features are obtained from a certain localized region in the image.

Hierarchical Question-Image Co-Attention for Visual Question Answering (NIPS 2016)

Question Features: Use 1-D convolution over word vectors (window=1,2,3). Concatenate these features and use LSTM to get $\mathbf{Q} = [q_1, \ldots, q_T] \in \mathbb{R}^{d \times T}$

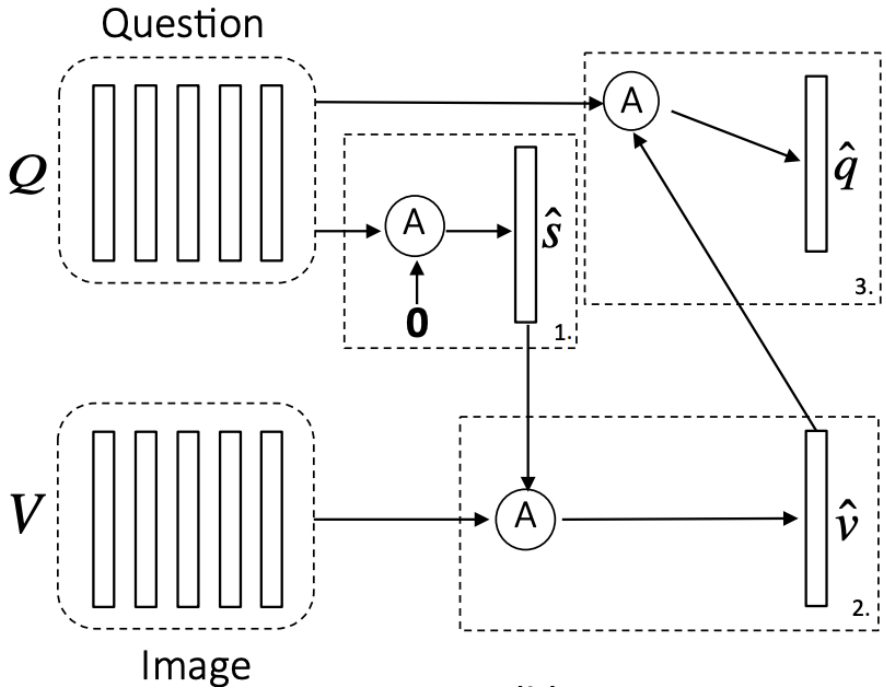
Image Features: Obtained from a pre-trained CNN on ImageNet (e.g., ResNet) $\mathbf{V} \in \mathbb{R}^{d \times N}$



Hierarchical Question-Image Co-Attention for Visual Question Answering (NIPS 2016)

Attention Operation: $\hat{\mathbf{x}} = \mathbb{A}(\mathbf{X}; \mathbf{g})$

X is the Image/Question and **g** is the attention guidance from Question/Image



$$\boldsymbol{H} = \tanh(\boldsymbol{W}_x \boldsymbol{X} + (\boldsymbol{W}_g \boldsymbol{g})\mathbb{1}^T)$$
$$\boldsymbol{a}^x = \text{softmax}(\boldsymbol{w}_{hx}^T \boldsymbol{H})$$
$$\hat{\boldsymbol{x}} = \sum a_i^x \boldsymbol{x}_i$$

$\mathbf{W}_x, \mathbf{W}_g \in \mathbb{R}^{k \times d}, \mathbf{w}_{hx} \in \mathbb{R}^k$

Initially, $\mathbf{X} = \mathbf{Q}, \mathbf{g} = 0$ and the output is $\hat{\mathbf{s}}$. Next, $\mathbf{X} = \mathbf{V}$ and $\mathbf{g} = \hat{\mathbf{s}}$ and the output is $\hat{\mathbf{v}}$. In the last step, again, $\mathbf{X} = \mathbf{Q}$ but $\mathbf{g} = \hat{\mathbf{v}}$.

Hierarchical Question-Image Co-Attention for Visual Question Answering (NIPS 2016)

Table 1: Results on the VQA dataset. "-" indicates the results is not available.

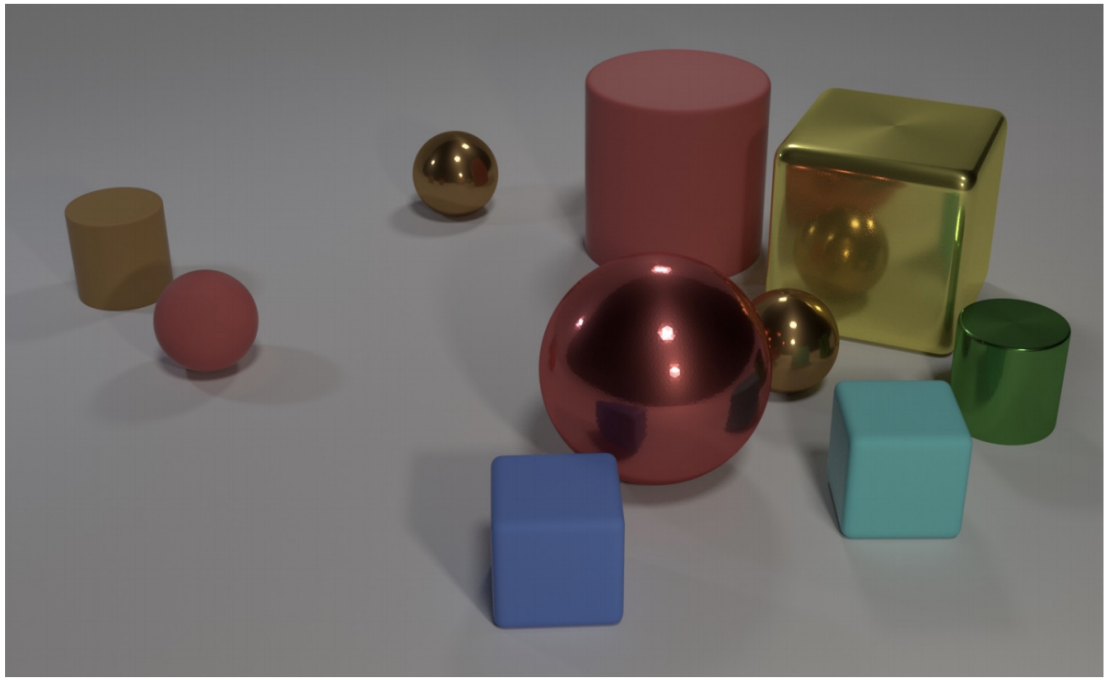
	Open-Ended					Multiple-Choice				
	test-dev				test-std	test-dev				test-std
Method	Y/N	Num	Other	All	All	Y/N	Num	Other	All	All
LSTM Q+I [2]	80.5	36.8	43.0	57.8	58.2	80.5	38.2	53.0	62.7	63.1
Region Sel. [20]	-	-	-	-	-	77.6	34.3	55.8	62.4	-
SMem [24]	80.9	37.3	43.1	58.0	58.2	-	-	-	-	-
SAN [25]	79.3	36.6	46.1	58.7	58.9	-	-	-	-	-
FDA [11]	**81.1**	36.2	45.8	59.2	59.5	**81.5**	39.0	54.7	64.0	64.2
DMN+ [23]	80.5	36.8	48.3	60.3	60.4	-	-	-	-	-
Ours^p+VGG	79.5	**38.7**	48.3	60.1	-	79.5	39.8	57.4	64.6	-
Ours^a+VGG	79.6	38.4	49.1	60.5	-	79.7	**40.1**	57.9	64.9	-
Ours^a+ResNet	79.7	**38.7**	**51.7**	**61.8**	**62.1**	79.7	40.0	**59.8**	**65.8**	**66.1**

Table 2: Results on the COCO-QA dataset. "-" indicates the results is not available.

Method	Object	Number	Color	Location	Accuracy	WUPS0.9	WUPS0.0
2-VIS+BLSTM [17]	58.2	44.8	49.5	47.3	55.1	65.3	88.6
IMG-CNN [15]	-	-	-	-	58.4	68.5	89.7
SAN(2, CNN) [25]	64.5	48.6	57.9	54.0	61.6	71.6	90.9
Ours^p+VGG	65.6	49.6	61.5	56.8	63.3	73.0	91.3
Ours^a+VGG	65.6	48.9	59.8	56.7	62.9	72.8	91.3
Ours^a+ResNet	**68.0**	**51.0**	**62.9**	**58.8**	**65.4**	**75.1**	**92.0**

CLEVR Dataset

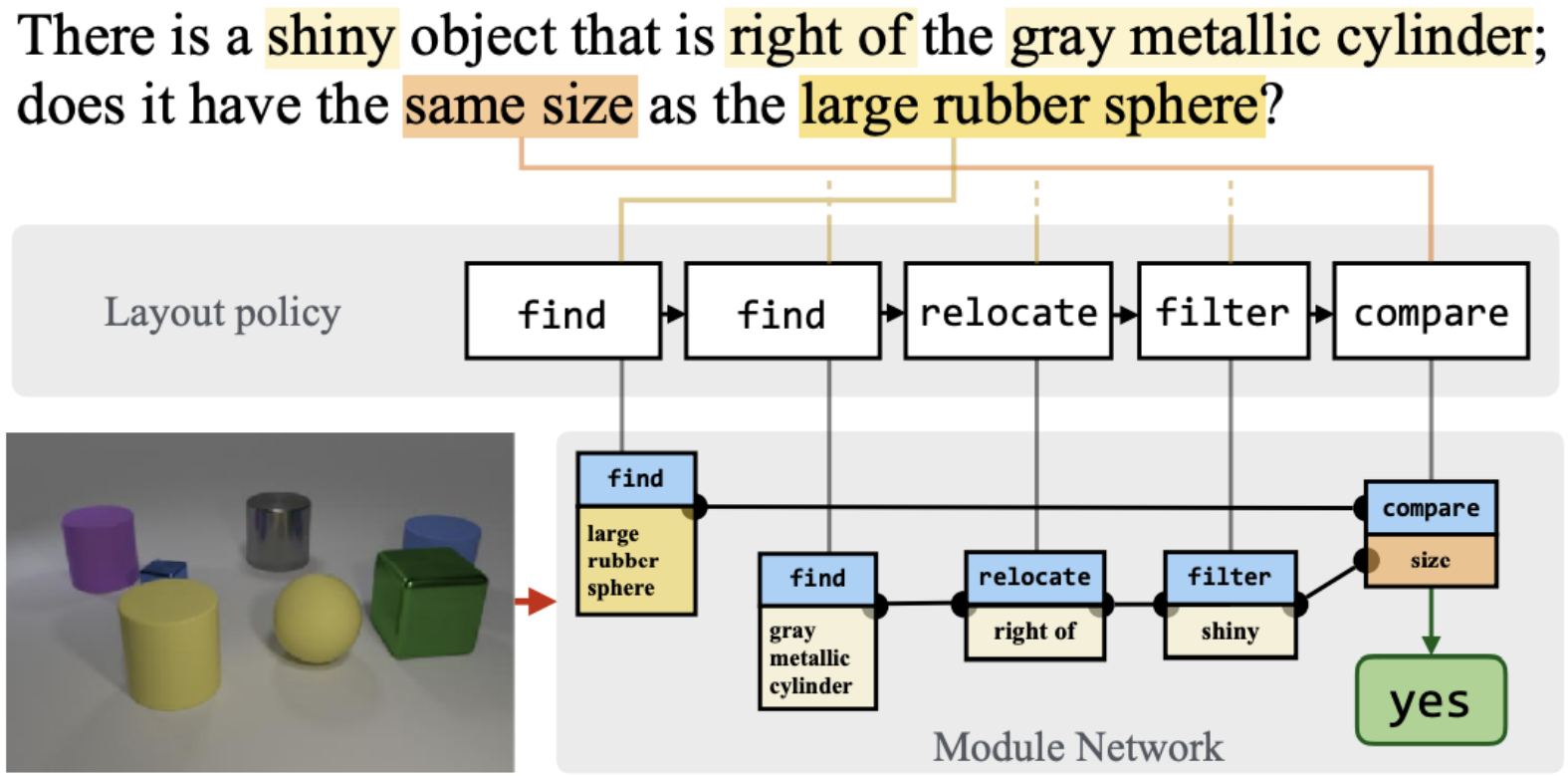
CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning (CVPR 2017)



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?
Q: How many objects are either small cylinders or metal things?

Learning to Reason: End-to-End Module Networks for Visual Question Answering (2017)

Idea: Translate question into a compositional structured query format
Motivation: Questions for VQA often ask for the location, count, size, *etc.* of some objects.



Learning to Reason: End-to-End Module Networks for Visual Question Answering (2017)

Fixed set of modules where each module performs some computation on the inputs given. The parameters in each module are learnable.
a_i's are attention maps over image and x_{vis}, x_{txt} are the features from image and question.

Module name	Att-inputs	Features	Output	Implementation details
`find`	(none)	x_{vis}, x_{txt}	att	$a_{out} = \text{conv}_2\left(\text{conv}_1(x_{vis}) \odot W x_{txt}\right)$
`relocate`	a	x_{vis}, x_{txt}	att	$a_{out} = \text{conv}_2\left(\text{conv}_1(x_{vis}) \odot W_1 \text{sum}(a \odot x_{vis}) \odot W_2 x_{txt}\right)$
`and`	a_1, a_2	(none)	att	$a_{out} = \text{minimum}(a_1, a_2)$
`or`	a_1, a_2	(none)	att	$a_{out} = \text{maximum}(a_1, a_2)$
`filter`	a	x_{vis}, x_{txt}	att	$a_{out} = \textbf{and}(a, \textbf{find}[x_{vis}, x_{txt}]())$, *i.e.* reusing `find` and `and`
`[exist, count]`	a	(none)	ans	$y = W^T \text{vec}(a)$
`describe`	a	x_{vis}, x_{txt}	ans	$y = W_1^T \left(W_2 \text{sum}(a \odot x_{vis}) \odot W_3 x_{txt}\right)$
`[eq_count, more, less]`	a_1, a_2	(none)	ans	$y = W_1^T \text{vec}(a_1) + W_2^T \text{vec}(a_2)$
`compare`	a_1, a_2	x_{vis}, x_{txt}	ans	$y = W_1^T \left(W_2 \text{sum}(a_1 \odot x_{vis}) \odot W_3 \text{sum}(a_2 \odot x_{vis}) \odot W_4 x_{txt}\right)$

Represent each question as:
$f_{m2}(f_{m4}(f_{m1}), f_{m3}(f_{m1}, f_{m1}))$

Learning to Reason: End-to-End Module Networks for Visual Question Answering (2017)

Policy: Given question, what should the layout be? *i.e.*, $p(l|q;\theta_{\text{layout}})$
"What object is next to the table?" $\rightarrow$ **describe(relocate(find()))**
θ_{layout} are the parameters of a Seq2Seq model.
Encoder: Given a question $[q_1, \ldots, q_T]$, use a multi-layer LSTM to encode the question as $[h_1, \ldots, h_T]$.
Decoder: At every step, predict a soft-attention map over the question words (α_{ti})

$$u_{ti} = v^T \tanh(W_1 h_i + W_2 h_t)$$

$$\alpha_{ti} = \frac{\exp(u_{ti})}{\sum_{j=1}^{T} \exp(u_{tj})}$$

$$x_{txt}^{(m)} = \sum_{i=1}^{T} \alpha_i^{(m)} w_i$$

$$c_t = \sum_{i=1}^{T} \alpha_{ti} h_i$$

$$p(l|q) = \prod_{m^{(t)} \in l} p(m^{(t)}|m^{(1)}, \ldots, m^{(t-1)}, q)$$

$$p(m^{(t)}|m^{(1)}, \cdots, m^{(t-1)}, q) = \text{softmax}(W_3 h_t + W_4 c_t)$$

Learning to Reason: End-to-End Module Networks for Visual Question Answering (2017)

Loss Function:

$$L(\theta) = E_{l \sim p(l|q;\theta)}[\tilde{L}(\theta, l; q, I)]$$

$\tilde{L}$ is the loss for a given network. $L(\theta)$ is not fully differentiable because l is discrete.

$\therefore$ Use backpropagation for differentiable parts and policy gradient for non-differentiable part

$$\nabla_\theta L = \mathbb{E}_{l \sim p(l|q;\theta)} \left[\tilde{L}(\theta, l) \nabla_\theta log(p(l_m|q;\theta) + \nabla_\theta \tilde{L}(\theta, l)) \right]$$

Can be estimated using Monte-Carlo Sampling

$$\nabla_\theta L \approx \frac{1}{M} \sum_{m=1}^{M} \left(\tilde{L}(\theta, l_m) \nabla_\theta \log p(l_m|q;\theta) + \nabla_\theta \tilde{L}(\theta, l_m) \right)$$

Learning to Reason: End-to-End Module Networks for Visual Question Answering (2017)

Reducing Variance:

$\tilde{L}(\theta, l_m) - b$

b is an exponential moving average over $\tilde{L}$

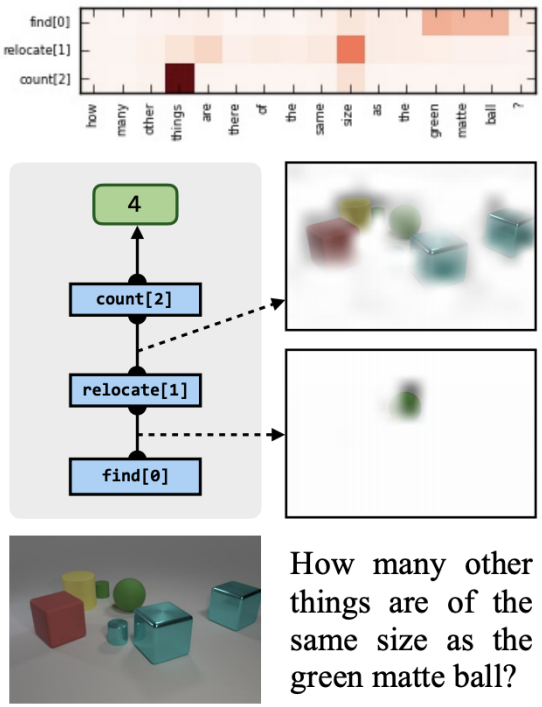
Behavior Cloning: Initialize policy to an expert policy which maybe sub-optimal. The expert policy is defined by manual rules.
For the initialization, the KL-Divergence between the expert policy p_e and p is minimized along with $\tilde{L}$ *i.e.*, $KL(p_e||p)$

Reason: Learning a policy from scratch is hard because there are multiple parameters (seq2seq LSTMs, attention weights, module parameters) to optimize.

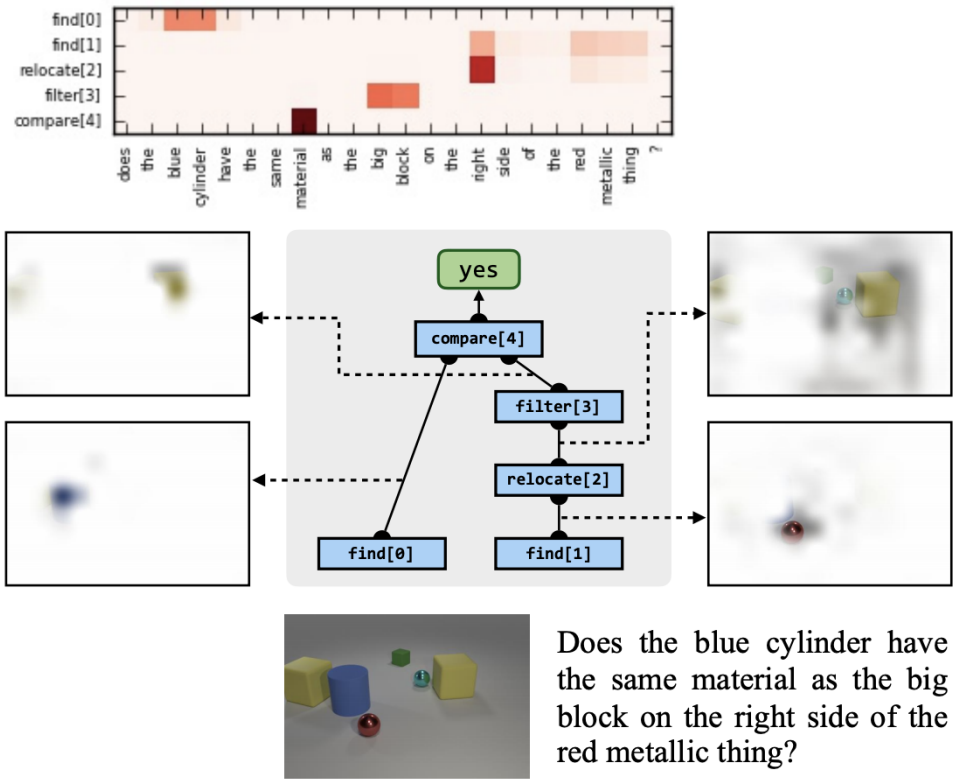
Learning to Reason: End-to-End Module Networks for Visual Question Answering (2017)

Method	**Overall**	Exist	Count	Compare Integer			Query Attribute				Compare Attribute			
				equal	less	more	size	color	material	shape	size	color	material	shape
CNN+BoW [26]	48.4	59.5	38.9	50	54	49	56	32	58	47	52	52	51	52
CNN+LSTM [4]	52.3	65.2	43.7	57	72	69	59	32	58	48	54	54	51	53
CNN+LSTM+MCB [9]	51.4	63.4	42.1	57	71	68	59	32	57	48	51	52	50	51
CNN+LSTM+SA [25]	68.5	71.1	52.2	60	82	74	87	81	88	85	52	55	51	51
NMN (expert layout) [3]	72.1	79.3	52.5	61.2	77.9	75.2	84.2	68.9	82.6	80.2	80.7	74.4	77.6	79.3
ours - policy search from scratch	69.0	72.7	55.1	71.6	85.1	79.0	88.1	74.0	86.6	84.1	50.1	53.9	48.6	51.1
ours - cloning expert	78.9	83.3	63.3	68.2	87.2	85.4	90.5	80.2	88.9	88.3	89.4	52.5	85.4	86.7
ours - policy search after cloning	**83.7**	**85.7**	**68.5**	**73.8**	**89.7**	**87.7**	**93.1**	**84.8**	**91.5**	**90.6**	**92.6**	**82.8**	**89.6**	**90.0**

Learning to Reason: End-to-End Module Networks for Visual Question Answering (2017)



How many other things are of the same size as the green matte ball?



Does the blue cylinder have the same material as the big block on the right side of the red metallic thing?