

Visual Question Answering

In Multimodal ML Lecture @ CMU

Rainer Stiefelhagen
Karlsruhe Institute of Technology

24.10.2019

Who am I / research / application domains



- Method-wise: no surprise, now almost everything deep-learning based
 - + research on multimodal interfaces / user studies / HCI

Current research topics

- How to deal with little (or no) training data
- How to model uncertainty
- How to cope with novelty

- Semi-/self supervised learning

- Vision and language
 - How to learn from text (wikipedia, movie plots, books, ...)

- Current applications / projects
 - Vision for the blind: Mobility & Orientation / Accessible documents
 - Medical image analysis
 - Driver monitoring (activities)
 - Movie analysis: semi-supervised person ID
 - Surveillance (person ID, person attributes)

About today's lecture

- (Updated) Part of my lecture on „Deep Learning for Computer Vision“

VL	Topic
1	Introduction / overview
2	Neural Networks - Basics
3	Deep CNN Networks Background
4	Deep Networks: Object Recognition
5	Scene Segmentation
6	Recurrent Neural Networks (RNN) + Embeddings
7	RNNs for image caption / tagging
8	Visual Question-Answering (VQA)
9	CNN learning in videos
10	Generative Adversarial Networks (GAN)
11	API / platforms / Tools / Summary

- You already know the basics CNNs, RNNS (LSTM/GRU), embeddings

About today's lecture: Image Analysis Tasks

■ classification

Q: what is the object?



■ localization

Q: where is X?



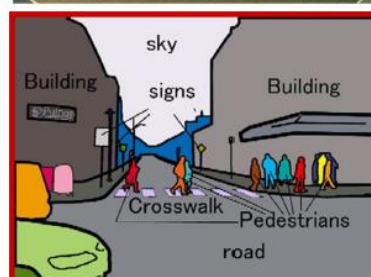
■ detection

Q: what and where?



■ segmentation

Q: what and where exactly?



difficulty
increases

About today's lecture: Image Analysis Tasks

Image & Video Captioning

A person riding a motorcycle on a dirt road.



Two dogs play in the grass.



A yellow school bus parked in a parking lot.

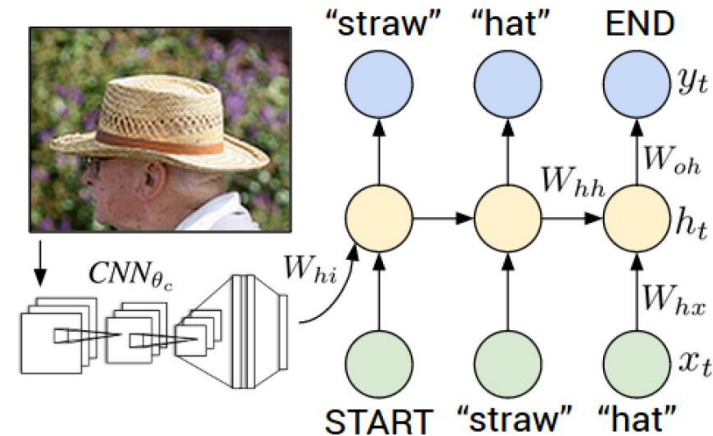


"black and white dog jumps over bar."



"young girl in pink shirt is swinging on swing."

Methods: CNN + RNN



CNNs + RNNs

Today's lecture: Visual Question Answering

- What is Visual Question Answering
- Data sets
- Approaches
 - Global embeddings
 - Using Attention / Stacked Attention Networks
 - Compositional Models
 - Memory Networks
 - Graph Networks
- VQA on Video

High-Level Understanding of Image Content: Visual and Video Question-Answering

- Image classification and information
 - Which object is in the center of the image?
 - What is the color of the ball?
- Action recognition
 - What are the boys playing?
 - What is the cat doing?
- Scene recognition
 - Where is the bus?
- Yes/No questions
 - Does this person have 20/20 vision?
- Counting
 - How many slices of pizza are there?



Different aspects of vision tested in QA

- Incorporates vision and language easily
- Questions can range in complexity
 - Simple visual tasks – classification, color, counting, etc.
 - Reasoning / induction – understand the world around us
- Ability to solve QA (of any type) will result in
 - More generic and stronger AI that can do everything!
 - e.g. can ask specific questions to search and rescue robots

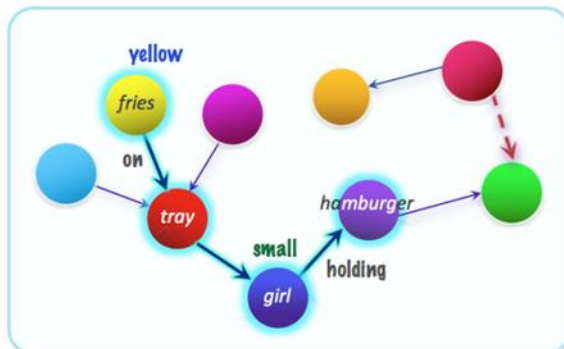
Question + Image → Answer

Overview of VQA Datasets

Visual Madlibs



GQA



VQA / VQA-v2

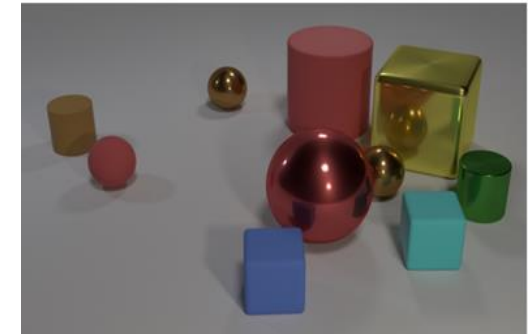


MovieQA



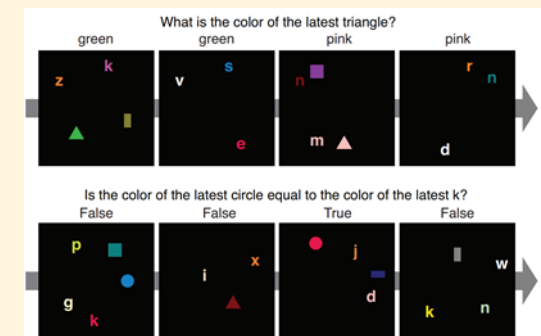
Synthetic Datasets

CLEVR



Are there an equal number of large things and metal spheres?

COG



Video Datasets

First Attempts – Visual Madlibs

- Fill in the blanks
- Based on COCO
- Crowd-sourced
- Multiple-choice
- 150k questions
- 12 types of questions



1. This place is a park.
2. When I look at this picture, I feel competitive.
3. The most interesting aspect of this picture is the guys playing shirtless.
4. One or two seconds before this picture was taken, the person caught the frisbee.
5. One or two seconds after this picture was taken, the guy will throw the frisbee.
6. Person A is wearing blue shorts.
7. Person A is in front of person B.
8. Person A is blocking person B.
9. Person B is a young man wearing an orange hat.
10. Person B is on a grassy field.
11. Person B is holding a frisbee.
12. The frisbee is white and round.
13. The frisbee is in the hand of the man with the orange cap.
14. People could throw the frisbee.
15. The people are playing with the frisbee.

Yu, et al. Visual Madlibs: Fill in the Blank Image Generation and Question Answering. In ICCV 2015.

First Attempts – Visual Madlibs

- image scene, emotion, interestingness
- image past, future
- object attribute, affordance, position
- person attribute, activity, location
- pairwise relationships

Type	Instruction	Prompt
1. image's scene	Describe the type of scene/place shown in this picture.	The place is a(n) ____ .
2. image's emotion	Describe the emotional content of this picture.	When I look at this picture, I feel ____ .
3. image's interesting	Describe the most interesting or unusual aspect of this picture.	The most interesting aspect of this picture is ____ .
4. image's past	Describe what happened immediately before this picture was taken.	One or two seconds before this picture was taken, ____ .
5. image's future	Describe what happened immediately after this picture was taken.	One or two seconds after this picture was taken, ____ .
6. object's attribute	Describe the appearance of the indicated object.	The object(s) is/are ____ .
7. object's affordance	Describe the function of the indicated object.	People could ____ the object(s).
8. object's position	Describe the position of the indicated object.	The object(s) is/are ____ .
9. person's attribute	Describe the appearance of the indicated person/people.	The person/people is/are ____ .
10. person's activity	Describe the activity of the indicated person/people.	The person/people is/are ____ .
11. person's location	Describe the location of the indicated person/people.	The person/people is/are ____ .
12. pair's relationship	Describe the relationship between the indicated person and object.	The person/people is/are ____ the object(s).

VQA dataset

- Based on the MS-COCO and Abstract Scenes datasets

- Massive crowd-sourcing
- 250K images
- 750K questions (v1)
- 1M questions (v2)

- Answering

- Multiple-choice
- Open-ended



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?

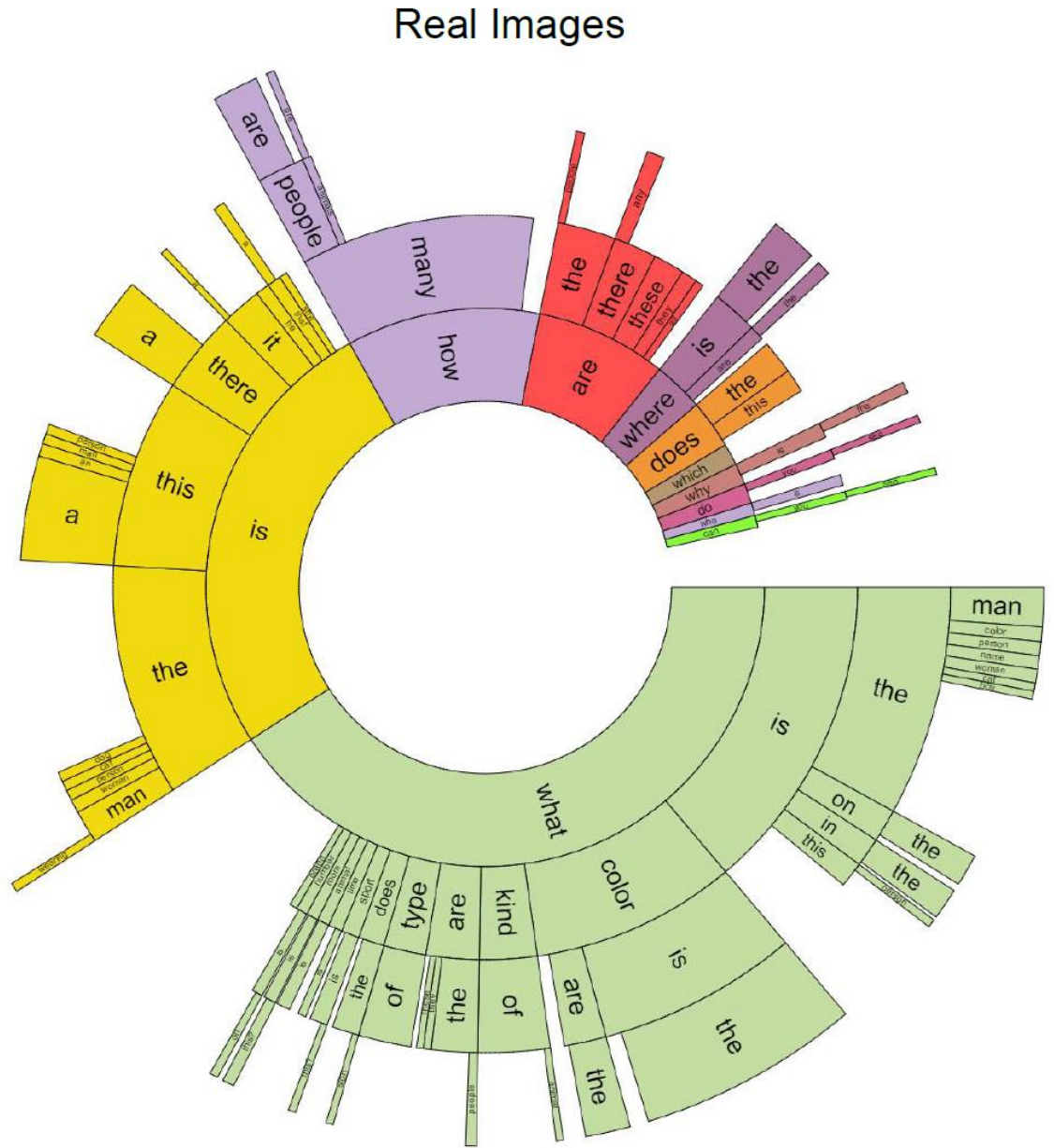


Does it appear to be rainy?
Does this person have 20/20 vision?

S. Antol, et al. VQA: Visual Question Answering. In ICCV 2015.

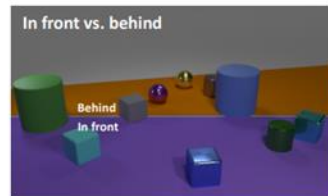
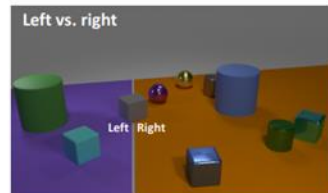
VQA dataset

- What is ...
 - What color ...
 - What kind / type ...
 - Is the ...
 - How many ...
 - Are there ...
 - Where is ...
 - Does the ...
-
- Yes/no
 - Objects, colors, counts

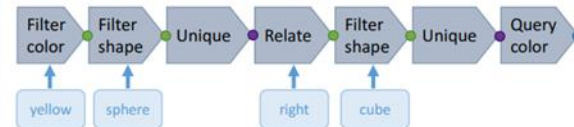


CLEVR

- Highly compositional questions on synthetic data
- Geometric forms in a 3D world
- Was able to generate 100K images and 1M questions
- Different types of questions: counting, existence, compare attributes

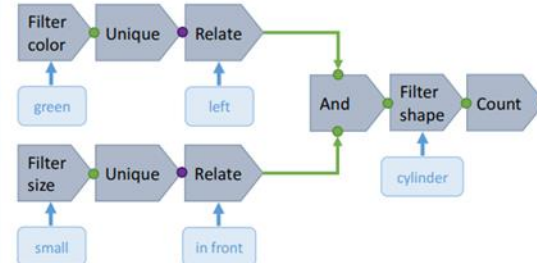


Sample chain-structured question:



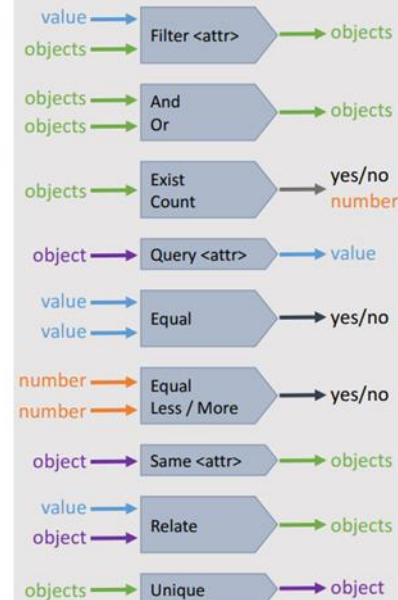
What color is the cube to the right of the yellow sphere?

Sample tree-structured question:



How many cylinders are in front of the small thing and on the left side of the green object?

CLEVR function catalog



Johnson et al., Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In CVPR 2017.

GQA Dataset for Compositional QA

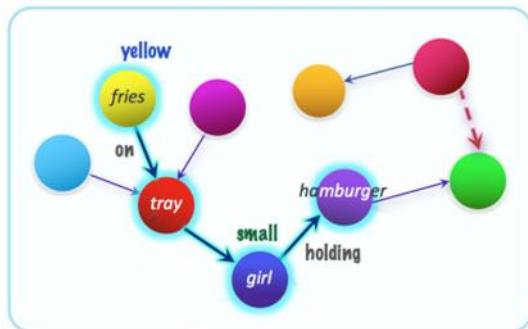
- Highly structured questions on *natural images*
- Questions generated using 524 templates
- Information from *scene graphs* used for producing the answer
- Around 22M questions in open-ended setting



Pattern: What|Which <type> [do you think] <is> <dobject>, <attr> or <decoy>?
Program: Select: <dobject> → Choose <type>: <attr>|<decoy>
Reference: The food on the red object left of the small girl that is holding a hamburger
Decoy: brown

What color is the food on the red object left of the small girl that is holding a hamburger, yellow or brown?

Select: hamburger → Relate: girl, holding → Filter size: small → Relate: object, left → Filter color: red → Relate: food, on → Choose color: yellow | brown



= Scene Graph Representation of the Image (manually annotated)

D. Hudson and C. Manning. GQA: a new dataset for compositional question answering over real-world images. CVPR 2019

VQA Types and Evaluation

■ Open-Ended

- Any answer possible
- Posed as a classification task
- Choose answer from answers found in the training set



Q: What is the mustache made of?
A: Bananas

■ Multiple-Choice

- Choose between multiple possible answers
- Each answer set additional to question as input to network

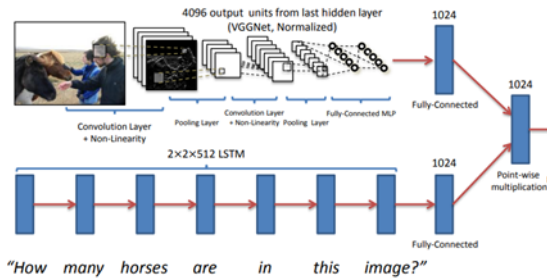


Q: What color is the tie?
A1: Red
A2: Blue
A3: Yellow

■ Evaluation Metric: Accuracy over all questions

Overview of Approaches for VQA

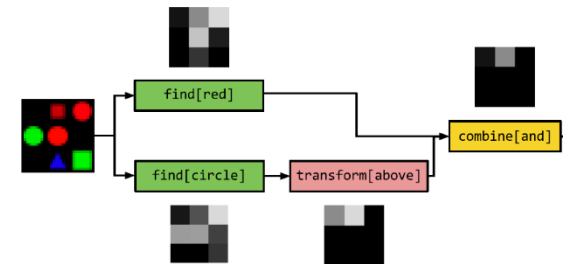
Global Embedding



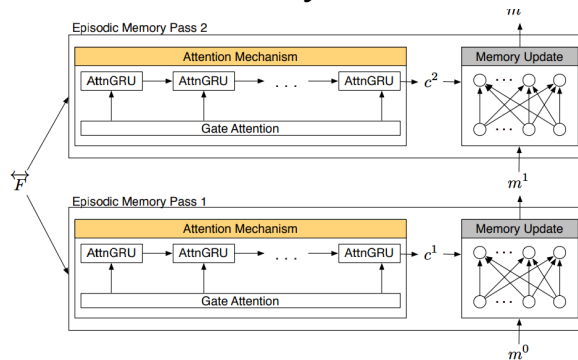
Attention-based



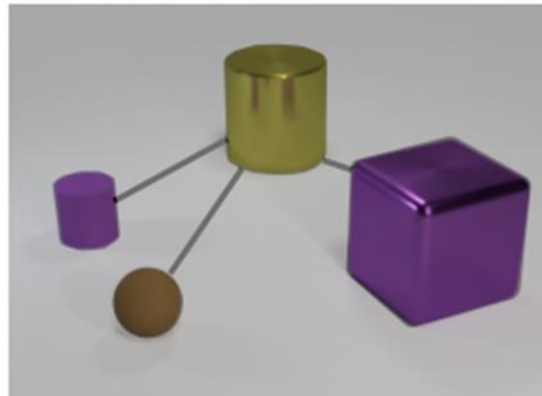
Compositional Models



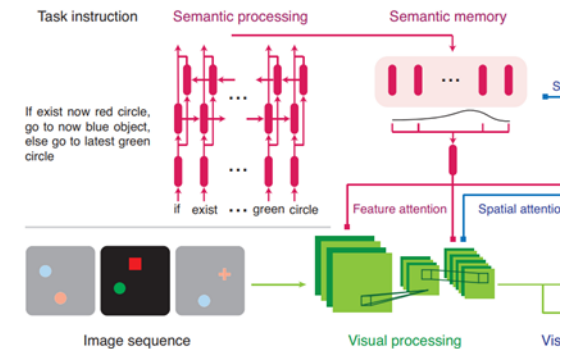
Memory Nets



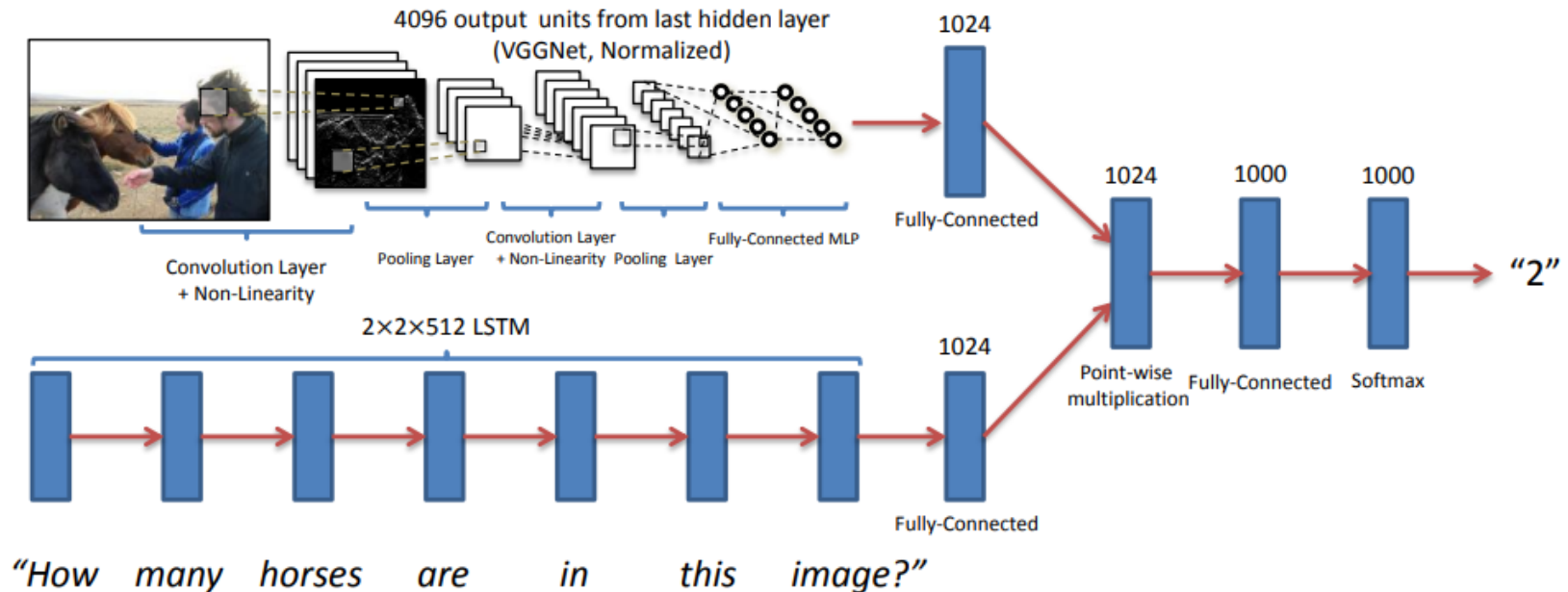
Graph Neural Networks



NNs for Videos



1. Global Embedding VQA Methods: CNN-RNN answering

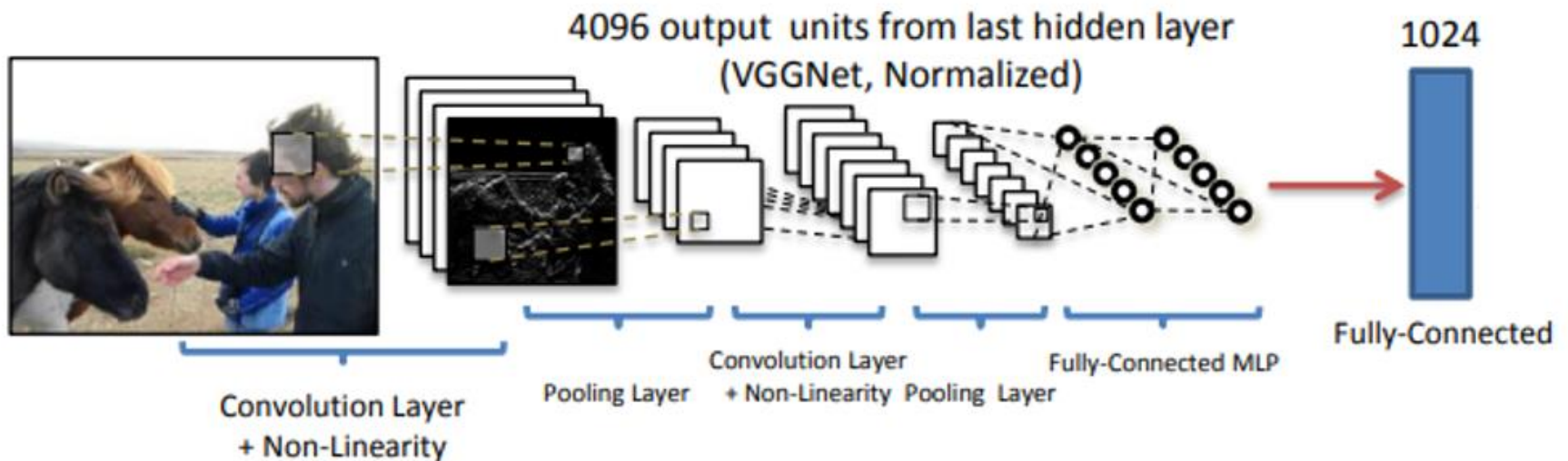


- Image embedded with CNN
- Question embedded using a LSTM
- Fusion / MLP to produce answer

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

1. Global Embedding VQA Methods: CNN-RNN answering

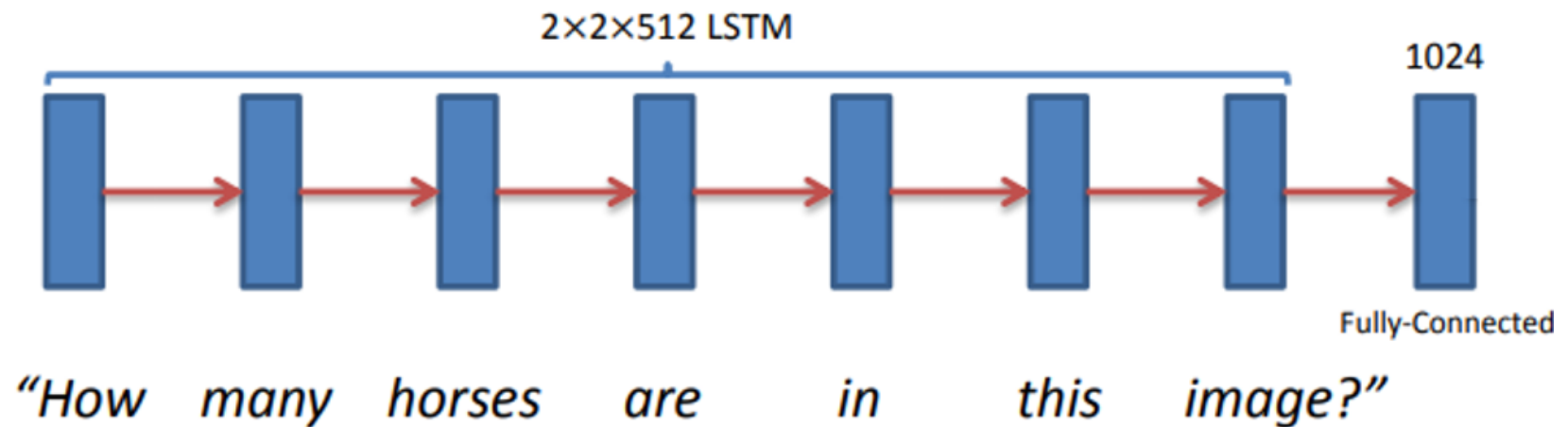
- L2 normalized activations from the last FC of VGG
- A second FC layer used to re-embed the visual features



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

1. Global Embedding VQA Methods: CNN-RNN answering

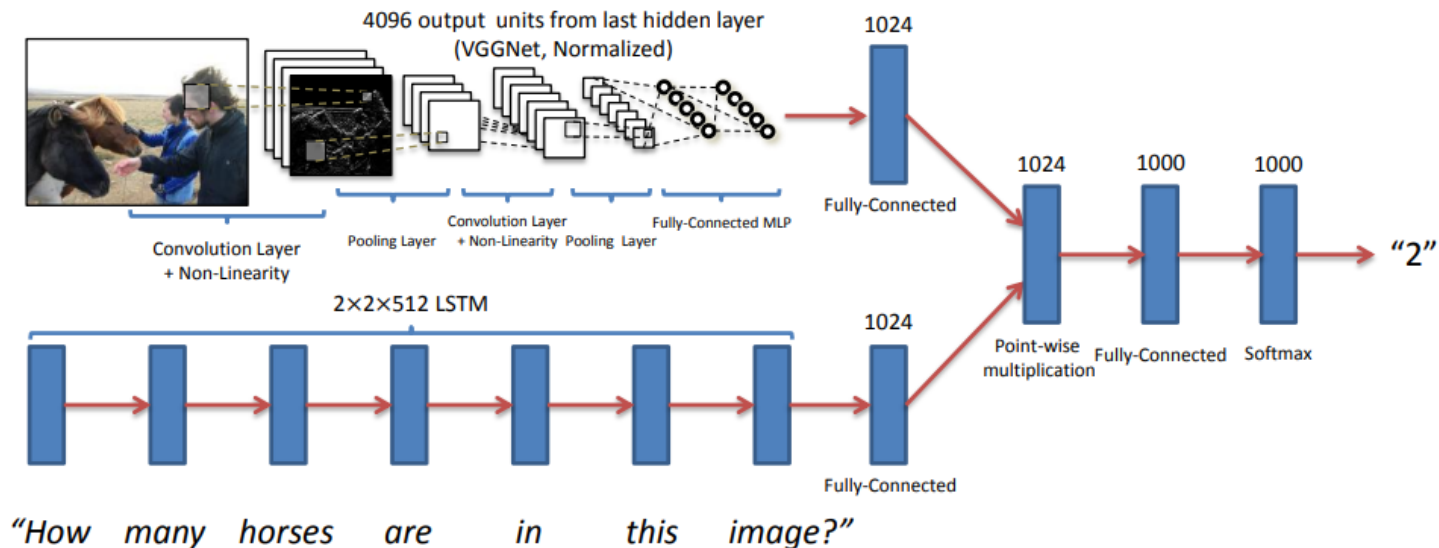
- Embedding of the question into a fixed vector representation
- 2-layered LSTM used on the words of the question
- Final representation = concatenation of last hidden and cell state



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

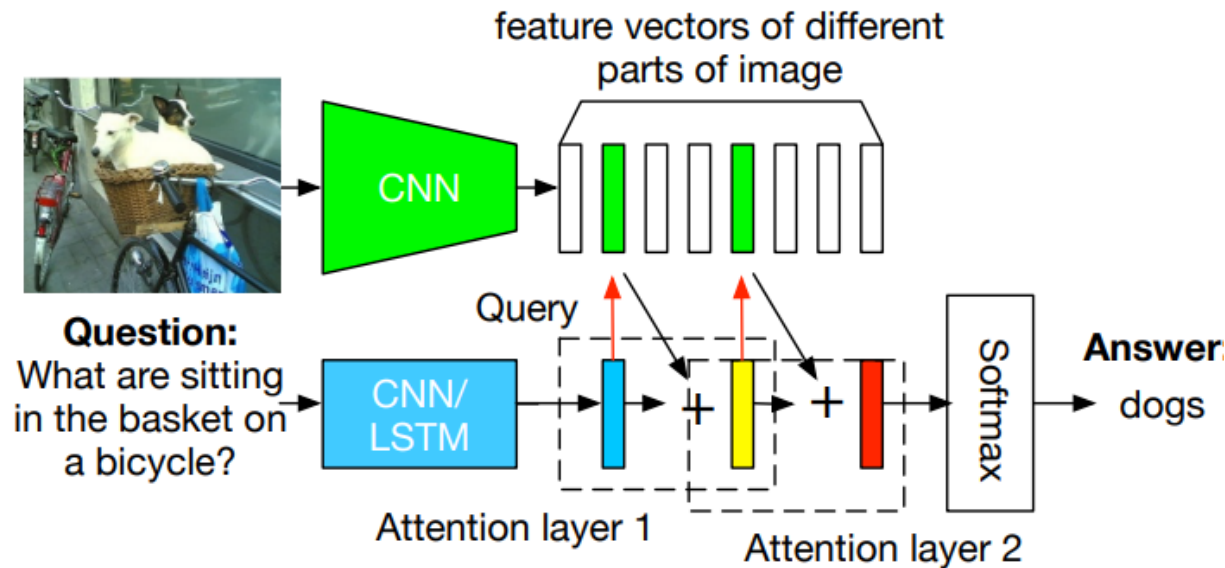
1. Global Embedding VQA Methods: CNN-RNN answering

- Next: Fusion of Question and Image Representation
- Multiple methods are possible: concatenation, addition
- Here: Multiplication (representations must have same dimensions)
- Fully connected layer with number of hidden units equal to number of possible answers generates the final prediction



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

2. Attention-based Networks: SAN – Stacked Attention Network

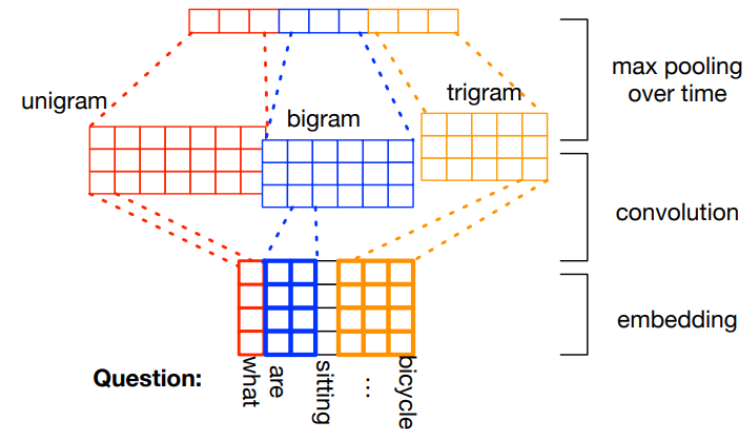


- Image model: CNN-based
- Question model: LSTM or CNN
- Stacked attention approach to focus on relevant parts and allow for multi-step reasoning

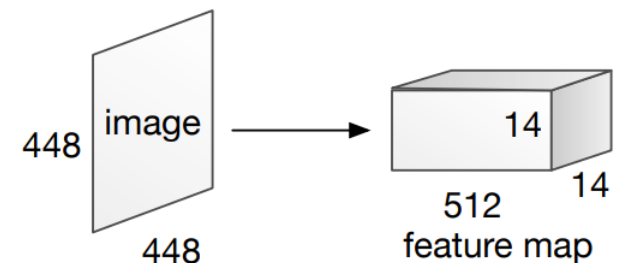
Yang et al. Stacked attention networks for image question answering. In CVPR 2016.

2. Attention-based Networks: SAN – Stacked Attention Network

- Question embedding:
 - 1D convolutions / filters of sizes 1, 2, 3
 - Maxpooling over time
 - → alternative to LSTMs



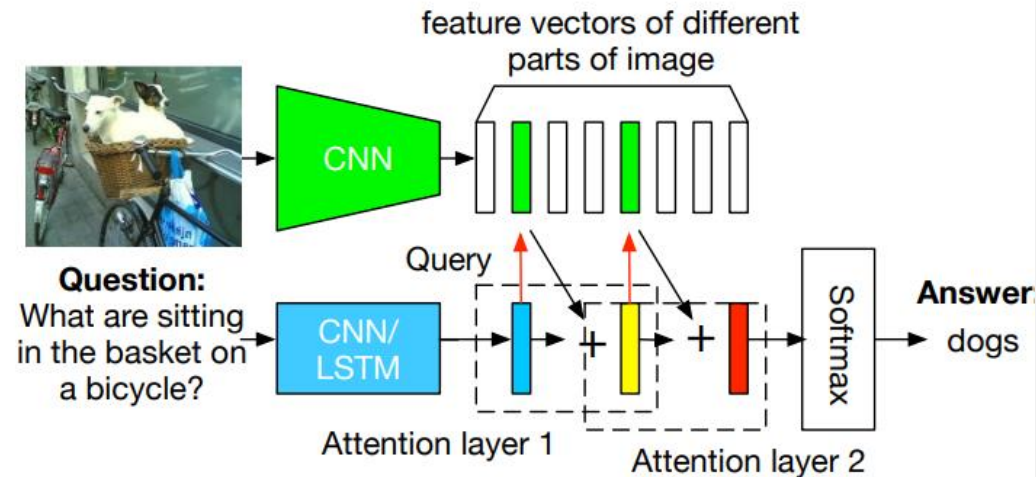
- Image embedding
 - Pre-trained VGG network
 - FC layers removed, obtaining a 3D tensor
 - Vector representation extracted for each image location



Yang et al. Stacked attention networks for image question answering. In CVPR 2016.

2. Attention-based Networks: SAN – Stacked Attention Network

- Uses the question to query image vectors in first layer
- Each query-visual vector pair gives out a confidence score / attention distribution (trained using single layer NN)
- Combines question vector and retrieved / weighted image vectors to form refined query
- Repeat N times (with new query embedding)
- Use Nth query embedding to answer question



(a) Stacked Attention Network for Image QA



Original Image

First Attention Layer

Second Attention Layer

Yang et al. Stacked attention networks for image question answering. In CVPR 2016.

2. Attention-based Networks: SAN – Stacked Attention Network

- Accuracy on the VQA dataset
- Strong improvement over non-attention-based methods

Methods	All	Yes/No	Number	Other
VQA: [1]				
Question	48.1	75.7	36.7	27.1
Image	28.1	64.0	0.4	3.8
Q+I	52.6	75.6	33.7	37.4
LSTM Q	48.8	78.2	35.7	26.6
LSTM Q+I	53.7	78.9	35.2	36.4
SAN(2, CNN)	58.7	79.3	36.6	46.1

Previously
discussed global
embedding scheme

Yang et al. Stacked attention networks for image question answering. In CVPR 2016.

2. SAN – Human Interpretability

- (a) What are pulling a man on a wagon down on dirt road?
Answer: horses Prediction: horses



- (b) What is the color of the box ?
Answer: red Prediction: red



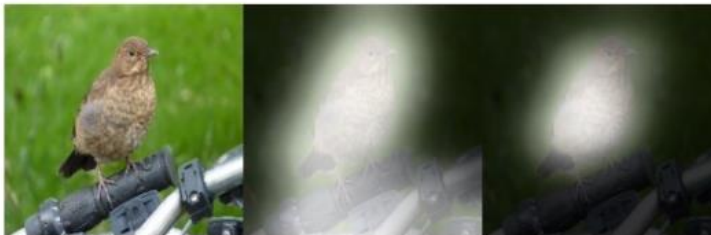
- (c) What next to the large umbrella attached to a table?
Answer: trees Prediction: tree



- (d) How many people are going up the mountain with walking sticks?
Answer: four Prediction: four



- (e) What is sitting on the handle bar of a bicycle?
Answer: bird Prediction: bird



- (f) What is the color of the horns?
Answer: red Prediction: red

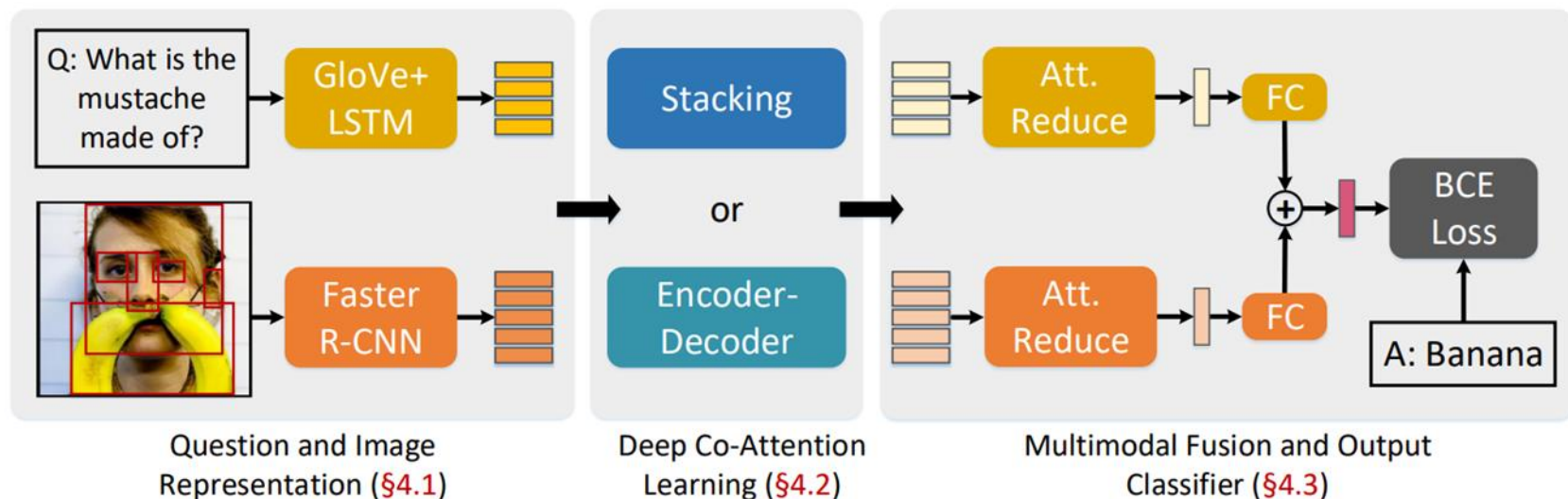


Original Image First Attention Layer Second Attention Layer Original Image First Attention Layer Second Attention Layer

Yang et al. Stacked attention networks for image question answering. In CVPR 2016.

2. Attention-based Networks: Trends (further reading)

- More focus on co-attention techniques
- Image represented as set of objects (localized by an obj. detector)
- Question tokenized into a set of words (GloVe word embeddings)
- Attend to words and image features simultaneously
- Dense interaction between modalities using stacking and enc-dec



Z. Yu, et al. Deep Modular Co-Attention Networks for Visual Question Answering. In *CVPR 2019*.

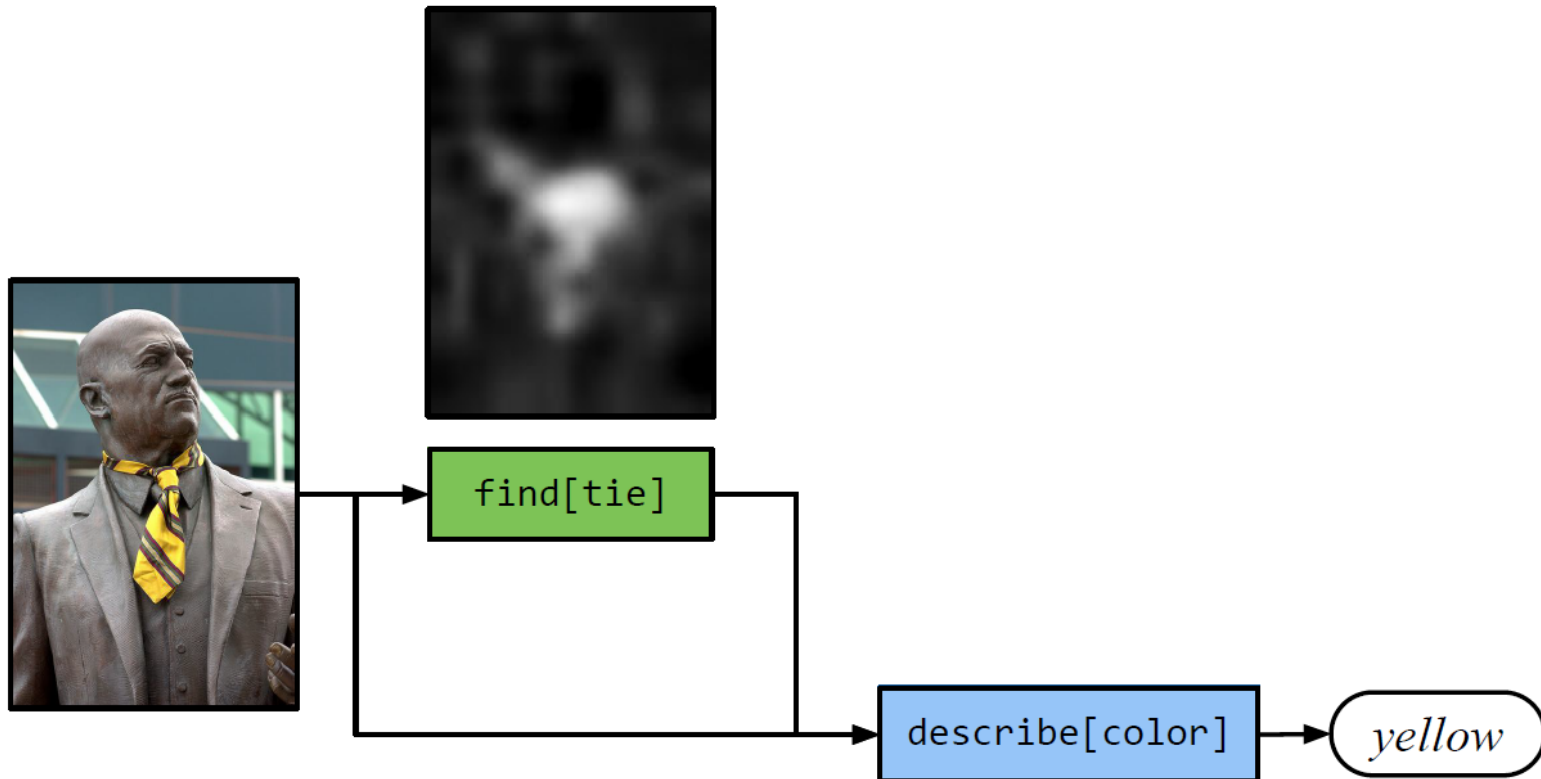
3. Compositional Models: Neural Module Networks

- Perform smaller/simpler tasks using specific neural networks
- Each training data as a 3-tuple – (question, image, answer)
- **Core idea:**
 - Decompose question into modular parts
 - Train special neural networks for these parts
- What color is the truck? → `color(truck)`
- Is there a circle next to the square? → `is(circle, next-to(square))`

J. Andreas, et al. Neural Module Networks. In *CVPR* 2016.

3. Compositional Models: NMN example of answering

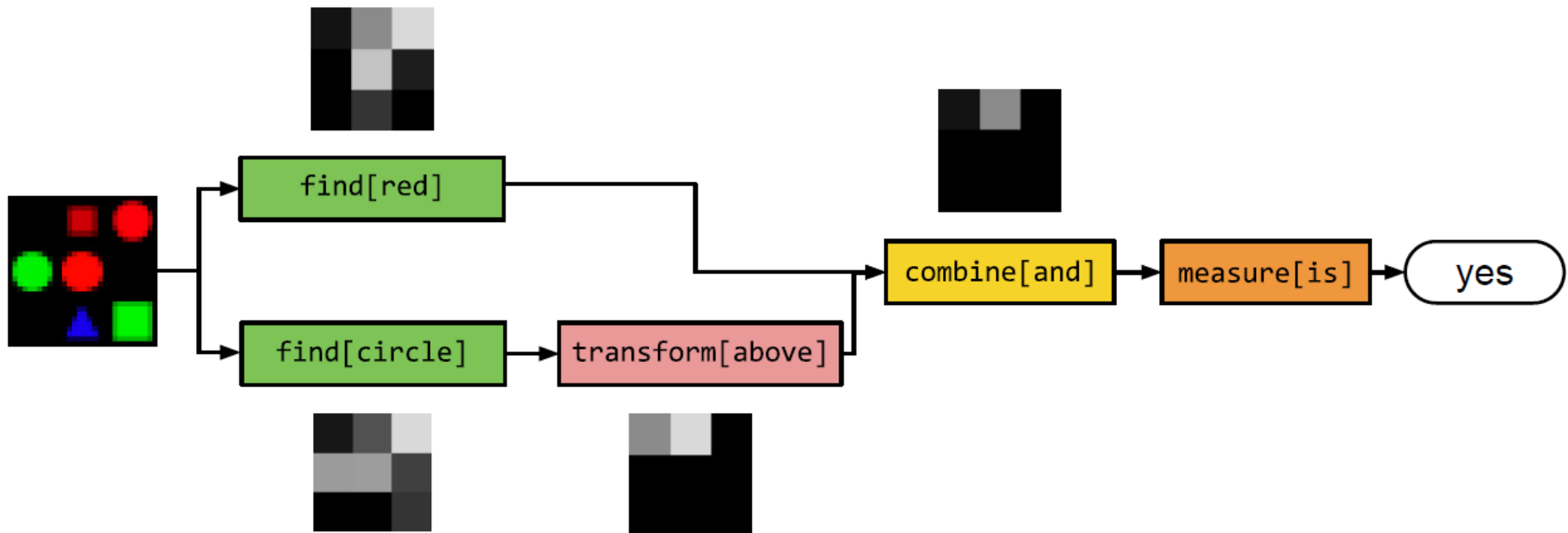
- Q: What color is his tie?
- Step 1. find the tie
- Step 2. describe it's color



J. Andreas, et al. Neural Module Networks. In *CVPR* 2016.

3. Compositional Models: NMN example of answering

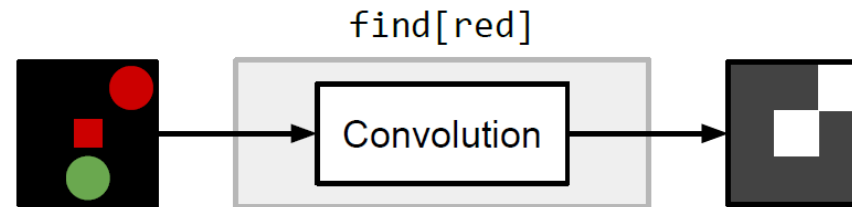
- Q. Is there a red shape above the circle
- Step 1. find red blobs
- Step 2. find circles, and search “above” the circle
- Step 3. combine and measure if it exists



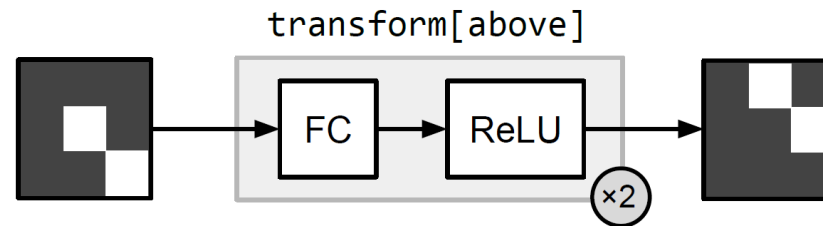
J. Andreas, et al. Neural Module Networks. In *CVPR* 2016.

3. Compositional Models: Modules (1)

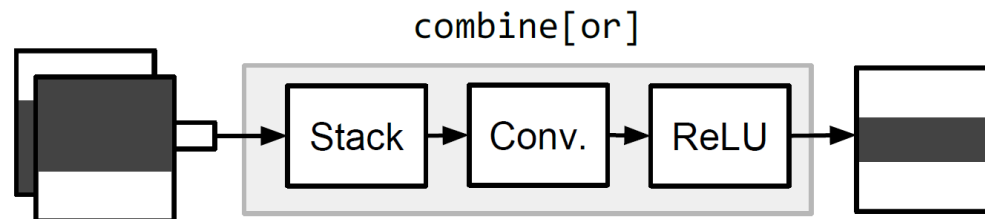
- Find: image \rightarrow attention



- Transform: attention \rightarrow attention



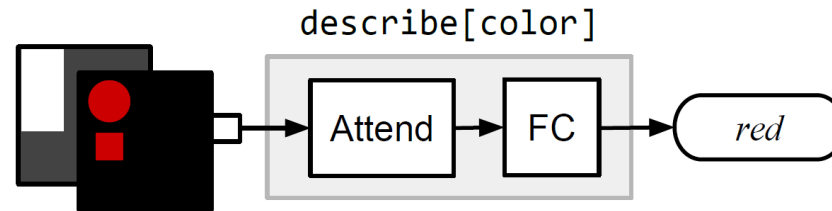
- Combine: attention X attention \rightarrow attention



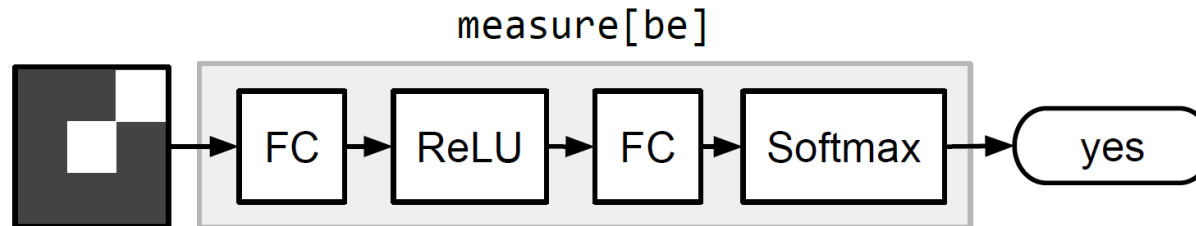
J. Andreas, et al. Neural Module Networks. In *CVPR* 2016.

3. Compositional Models: Modules (2)

- Describe: image X attention \rightarrow label



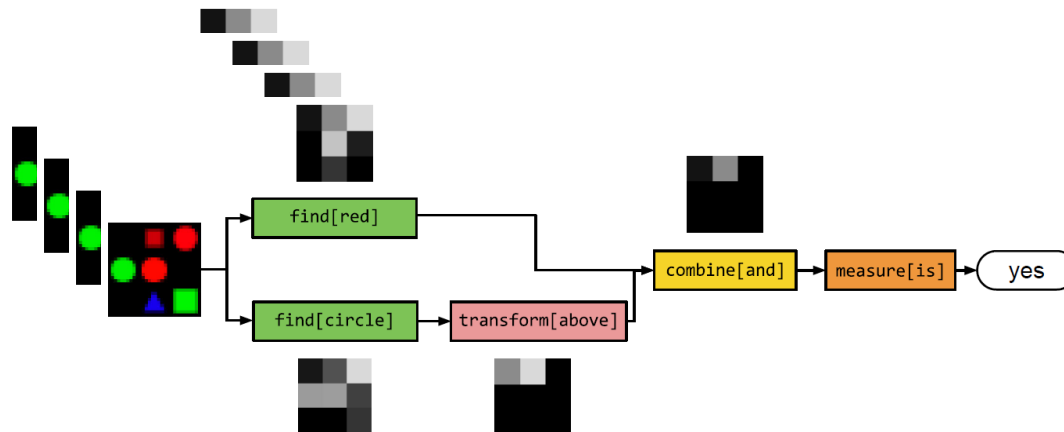
- Measure: attention \rightarrow label



J. Andreas, et al. Neural Module Networks. In *CVPR* 2016.

3. Compositional Models: Training NMN

- For each question in training dataset, build the neural module chains
- Question processed using parser
- Training the chain: supervision comes only from answer
- Share weights between similar functions and arguments
 - e.g. `find(red)` in all examples has same parameters
 - and `find(blue)` has different parameters, but same architecture



J. Andreas, et al. Neural Module Networks. In *CVPR* 2016.

3. Compositional Models: Results





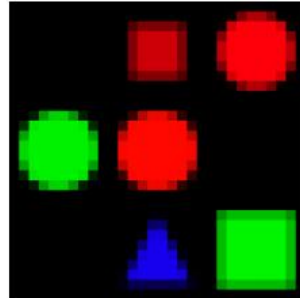
- Only small performance improvements (Accuracy on VQA)

	test-dev				test
	Yes/No	Number	Other	All	All
LSTM	78.7	36.6	28.1	49.8	—
VIS+LSTM [3] ²	78.9	35.2	36.4	53.7	54.1
ATT+LSTM	80.6	36.4	42.0	57.2	—
NMN	70.7	36.8	39.2	54.8	—
NMN+LSTM	81.2	35.2	43.3	58.0	—
NMN+LSTM+FT	81.2	38.0	44.0	58.6	58.7

J. Andreas, et al. Neural Module Networks. In *CVPR* 2016.

3. Compositional Models: Interpretable

- Only small performance improvements (Accuracy on VQA)
- However, model human interpretable

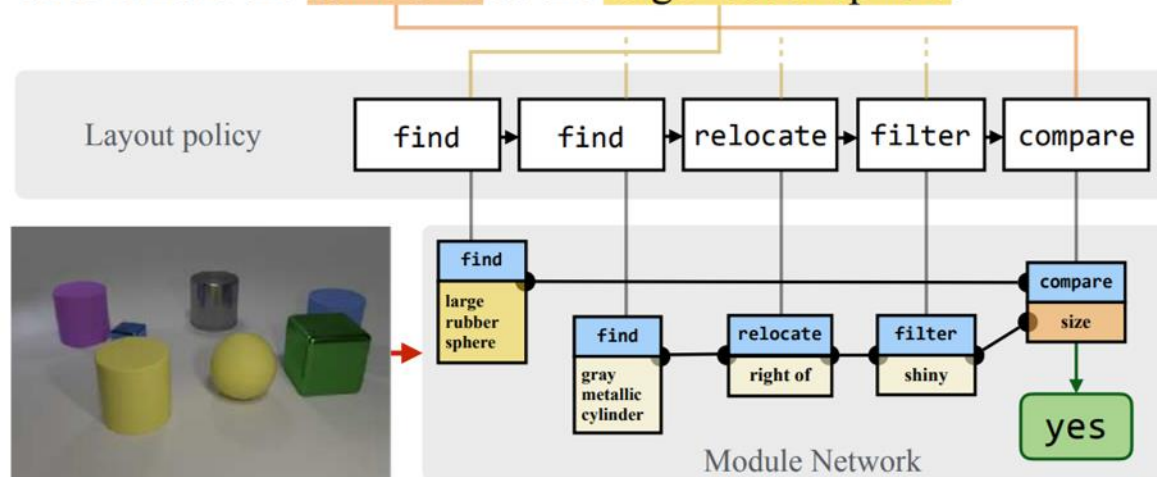
				
<i>how many different lights in various different shapes and sizes?</i>	<i>what is the color of the horse?</i>	<i>what color is the vase?</i>	<i>is the bus full of passengers?</i>	<i>is there a red shape above a circle?</i>
<code>measure[count](attend[light])</code>	<code>classify[color](attend[horse])</code>	<code>classify[color](attend[vase])</code>	<code>measure[is](combine[and](attend[bus], attend[full])</code>	<code>measure[is](combine[and](attend[red], re-attend[above](attend[circle]))))</code>
four (four)	brown (brown)	green (green)	yes (yes)	no (no)

J. Andreas, et al. Neural Module Networks. In *CVPR* 2016.

3. Compositional Models: Better – Train network *entirely* end-to-end

- NMN relies on parsers (which are not learned from the data)
- Here: model learns the layout directly using only QA pairs
- No separate instantiations with different parameters, but soft assignment
- However, layout generation discrete -> no full backpropagation
- Employs reinforcement learning techniques for training

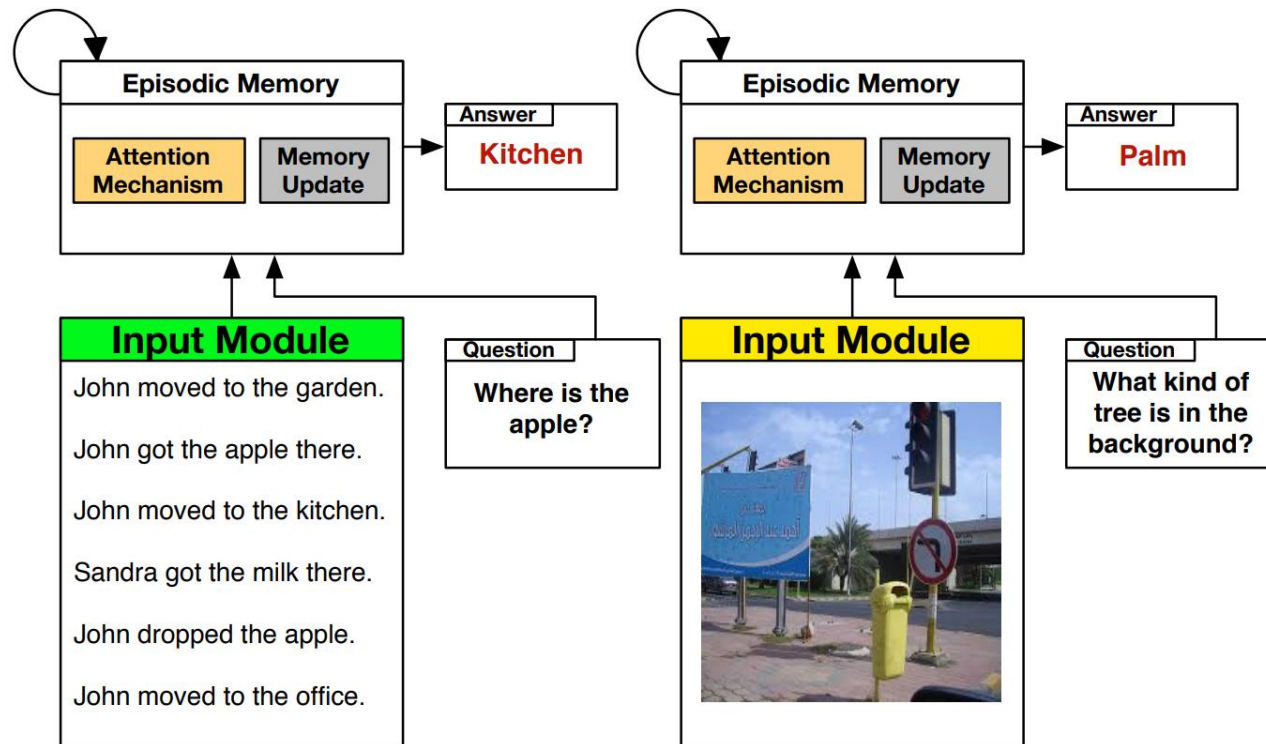
There is a shiny object that is right of the gray metallic cylinder;
does it have the same size as the large rubber sphere?



Hu, et al. Learning to reason: End-to-end module networks for visual question answering. In *ICCV* 2017.

4. Dynamic Memory Networks (DMN)

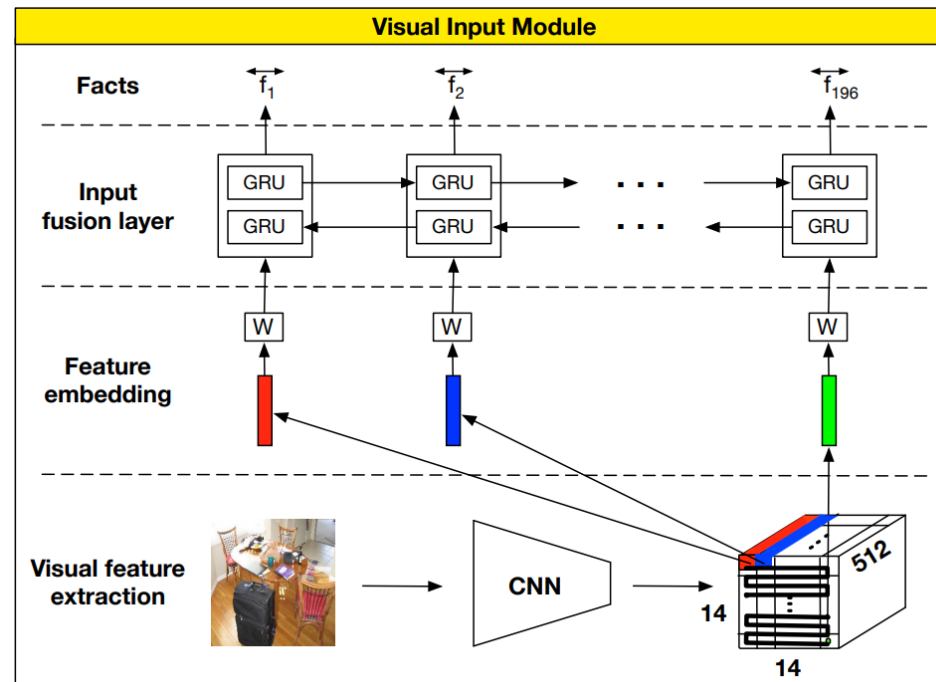
- Inspired by computer systems
- Reading and writing from a working memory
- Used for text and visual question answering



Xiong, et al. Dynamic Memory Networks for Visual and Textual Question Answering. In *ICML* 2016.

4. Dynamic Memory Networks (DMN)

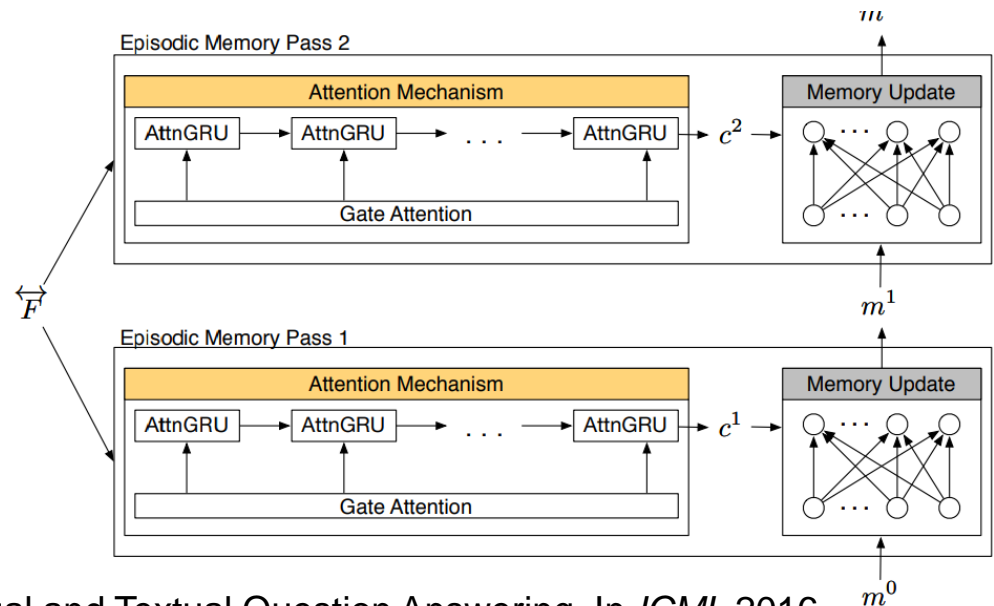
- Knowledge input: ordered set of facts (eg, sentences or img features)
- The set of facts are re-embedded using a GRU
- The hidden state is used as the embedding of the fact
- Question represented as the last hidden state of a second GRU (not shown here)



Xiong, et al. Dynamic Memory Networks for Visual and Textual Question Answering. In *ICML* 2016.

4. Dynamic Memory Networks (DMN)

- Memory (m_0) initialized with the question representation
- In each step the model reads and writes to the memory
- Read: using soft attention to extract information from memory = c^t
 - Attention values learned from facts (f_i), question and memory states
- Write: function on question and information extracted from memory (c^t)
- Final output of the memory is concatenated with the question for answer

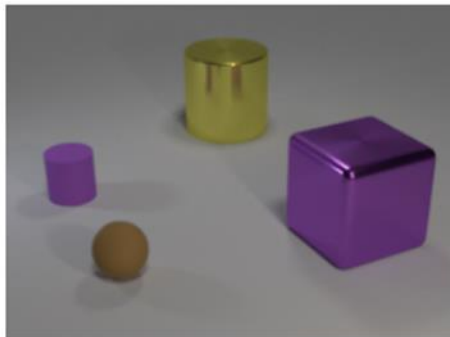


Xiong, et al. Dynamic Memory Networks for Visual and Textual Question Answering. In *ICML 2016*.

5. Graph Neural Networks: Relational Network

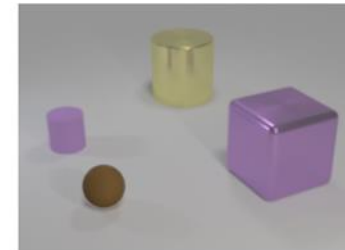
- Non-relational Questions usually *easy*
- Relational questions:
 - can get very complex
 - need multiple reasoning steps

Original Image:



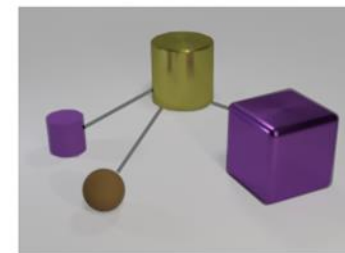
Non-relational question:

What is the size of the brown sphere?



Relational question:

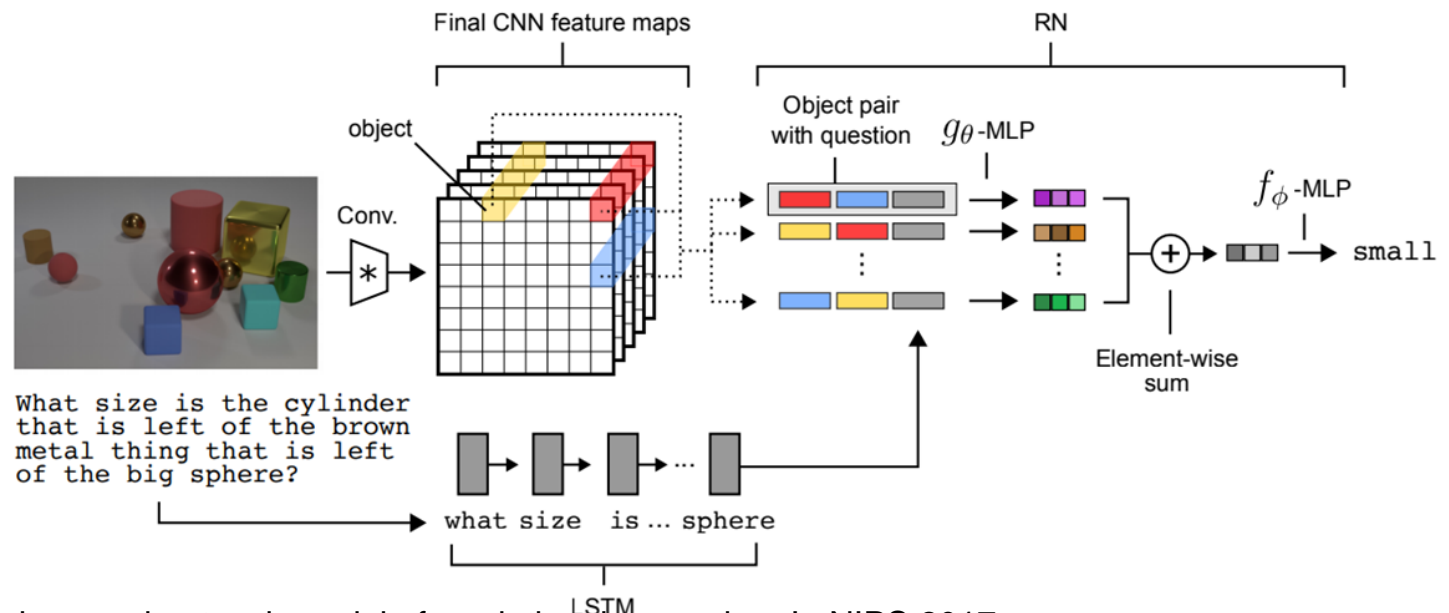
Are there any rubber things that have the same size as the yellow metallic cylinder?



Santoro, et al. A simple neural network module for relational reasoning. In NIPS 2017.

5. Graph Neural Networks: Relational Network

- A NN with a structure primed for relational reasoning
- Considers all pairs of objects (conditioned on the question)
- Image of size 128x128 encoded with a 4 layered CNN
- Each cell pair in the 3D tensor is concatenated with the question
- These representations are re-embedded using an MLP
- Prediction module takes the sum of this set of vectors
- Using a final FC layer an answer is generated



Santoro, et al. A simple neural network module for relational reasoning. In NIPS 2017.

5. Graph Neural Networks: Relational Network

- Exceeds human performance on CLEVR!
- Around 20% improvement over previous methods

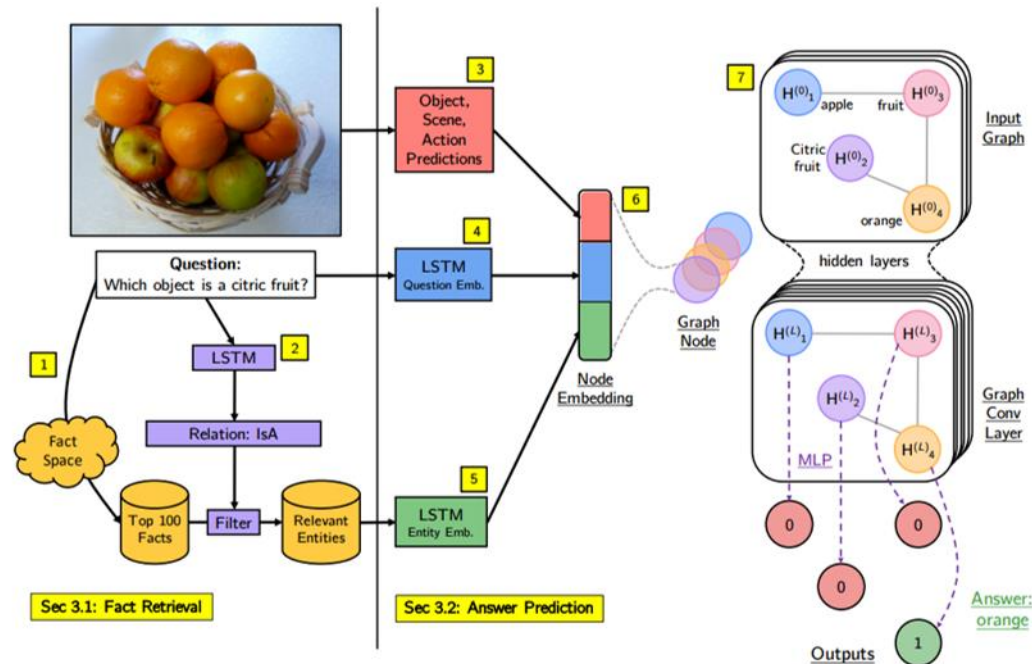
Model	Overall	Count	Exist	Compare Numbers	Query Attribute	Compare Attribute
Human	92.6	86.7	96.6	86.5	95.0	96.0
Q-type baseline	41.8	34.6	50.2	51.0	36.0	51.3
LSTM	46.8	41.7	61.1	69.8	36.8	51.8
CNN+LSTM	52.3	43.7	65.2	67.1	49.3	53.0
CNN+LSTM+SA	68.5	52.2	71.1	73.5	85.3	52.3
CNN+LSTM+SA*	76.6	64.4	82.7	77.4	82.6	75.4
CNN+LSTM+RN	95.5	90.1	97.8	93.6	97.9	97.1

Santoro, et al. A simple neural network module for relational reasoning. In NIPS 2017.

5. Graph Neural Networks: Trends (1)

Node Refinement Techniques: e.g., based on Graph Convolutions

- In each step refine the nodes (detected objects) using their neighbors
- Neighbors represented using adjacency matrix
- Adjacency matrix calculated (often) end-to-end

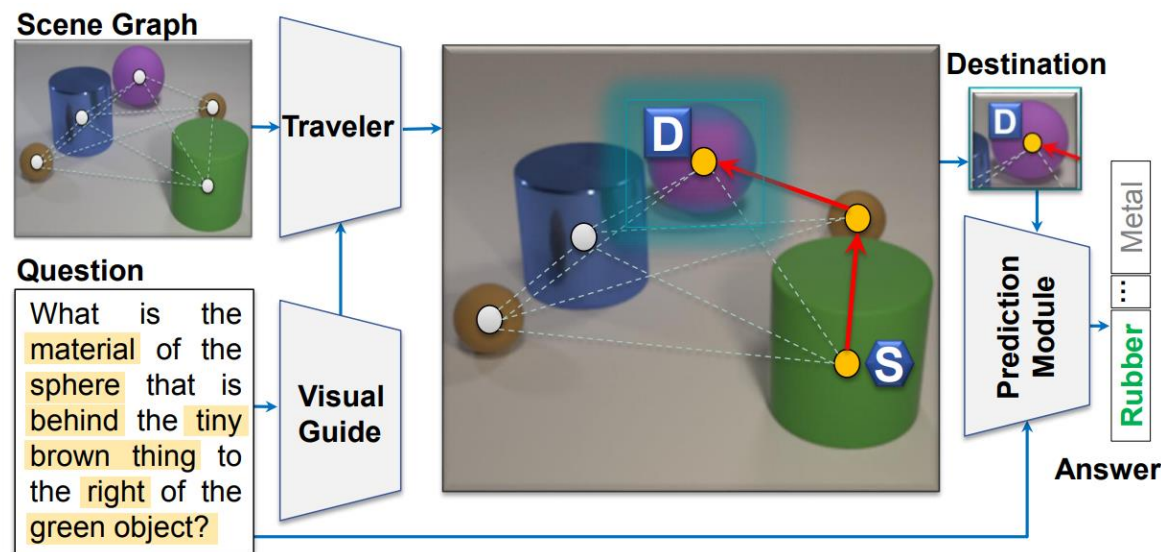


Narasimhan, et al. Out of the box: Reasoning with graph convolution nets for factual visual question answering. In NIPS 2018.

5. Graph Neural Networks: Trends (2)

Path-based approaches:

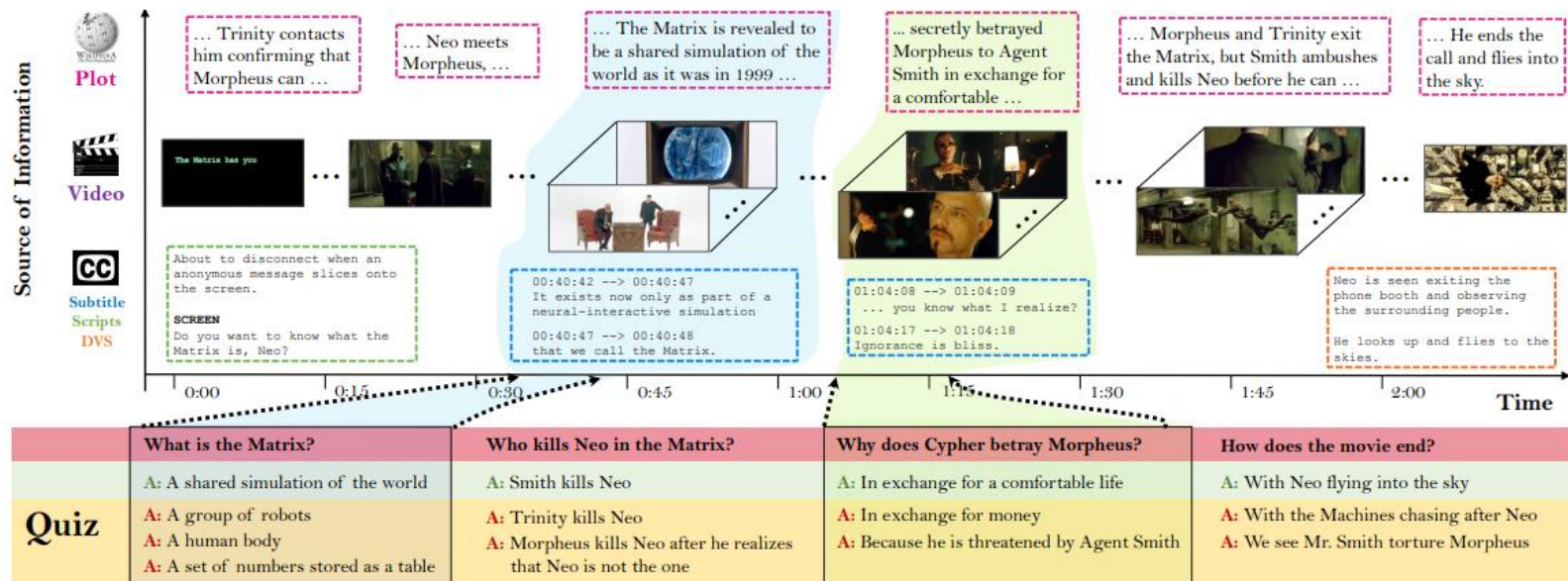
- Learns to follow paths *internally* using only QA as supervision
- Path described by multiple adjacency matrices
- No Node/Edge refinement during traversal
- The destination (alone) used with the question to generate answer



Haurilet et al. It's not about the Journey; It's about the Destination: Following Soft Paths under Question-Guidance for Visual Reasoning. In CVPR 2019.

6. Video Question Answering

- Videos are more complex than images
- Videos that tell a story are even more so
- Strongly dependence between frames
- Models have to choose relevant set of frames to answer question



MovieQA Dataset

- Our work on Story Question-Answering!
- Understand visual aspects: “who did what to whom and where”
- Also look at reasoning: “why and how” events took place
- Look at temporal reasoning: “what happens when ...”
- Need semantics to understand questions
 - e.g. How does the movie end?

Tapaswi, et al. MovieQA: Understanding stories in movies through question-answering. In CVPR. 2016.

MovieQA:

Answering with multiple story sources

Q. Who makes Indy return the crucifix after escaping from the grave robbers?

A. The local sheriff

VIDEO CLIP



PLOT

The men give chase, leaving Indy with a bloody cut across his chin from a bullwhip and a new phobia of snakes. Indy escapes, but the local **sheriff** makes him return the **crucifix**.

DVS

Herman means the Sheriff, who now enters the house. Indy shows the **Cross**, more or less handing it to the **Sheriff** to make his point. The **Sheriff** takes it casually.

SUBTITLE

00:10:50 --> 00:10:52
You still got it?
00:10:52 --> 00:10:53
Well, yes, sir.
00:10:53 --> 00:10:54
It's right here.
00:10:55 --> 00:10:56
I'm glad to see that
00:10:56 --> 00:10:59
because the rightful owner of this **cross**

SCRIPT

SHERIFF: You still got it?
INDY: Well, yes sir.
INDY: It's right here!
Indy shows the **CROSS**, more or less handing it to the **SHERIFF** to make his point. The **Sheriff** takes it casually.
SHERIFF: I'm glad to see that because the rightful owner of this **Cross** won't press ...

MovieQA examples

Titanic

How does Jack meet his end?



Freezes to death

Drowns as he can't swim

He is rescued but dies from frostbite

Dies as an old man in bed

Dies later in his life after obstacles

Indiana Jones: Last Crusade

What does Indy do to the grave robbers in the beginning?



He kills them while they are asleep

He steals their crucifix

He steals their horses

He tells the Boy Scouts to beat them

He calls the museum about the crucifix

MovieQA examples

LOTR: Return of the King

Why does Arwen wish to stay in Middle Earth?



She is too weak to travel

She wants to die on Middle Earth

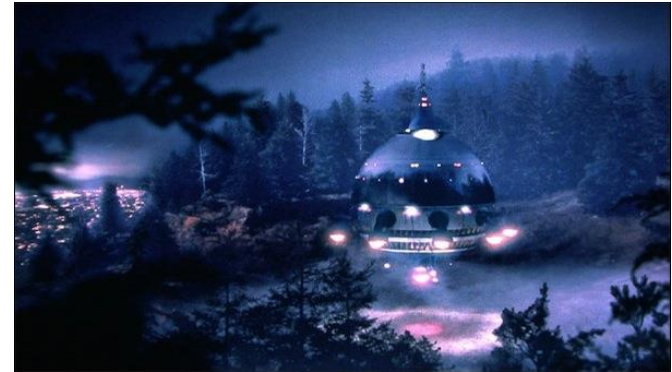
Her son asked her to stay

Arwen sees her son in her visions

She likes Middle Earth

E. T. the Extra-Terrestrial

Do aliens leave one of their own on Earth on purpose?



Yes, they leave it on purpose

No, they leave it accidentally

No, it falls off the spaceship

Yes, they leave it as a spy

They don't leave anyone

MovieQA examples

The Firm

Who tells Mitch about corrupt business of the Firm?

PLOT SYNOPSIS:

Mitch realizes he is trapped after two associates of the Firm die mysteriously. He is approached by **FBI agents** who inform him that that BL&L's biggest client is the Morolto Mafia family.

The Firm's senior partners

Mitch's coworkers

The FBI

The Moroltos

One of the Firm's clients

The Client

What are Mark and Ricky doing in the woods?

SUBTITLES:

- I wish. Sit here.
- **Don't try to swallow the smoke yet.**
- You're not ready for that.
- You'll just choke and puke all over the place.
- Suck a little and blow.

They are cutting woods

The are shooting each other

They are trying to kill each other

They are smoking cigarettes

They are having sex

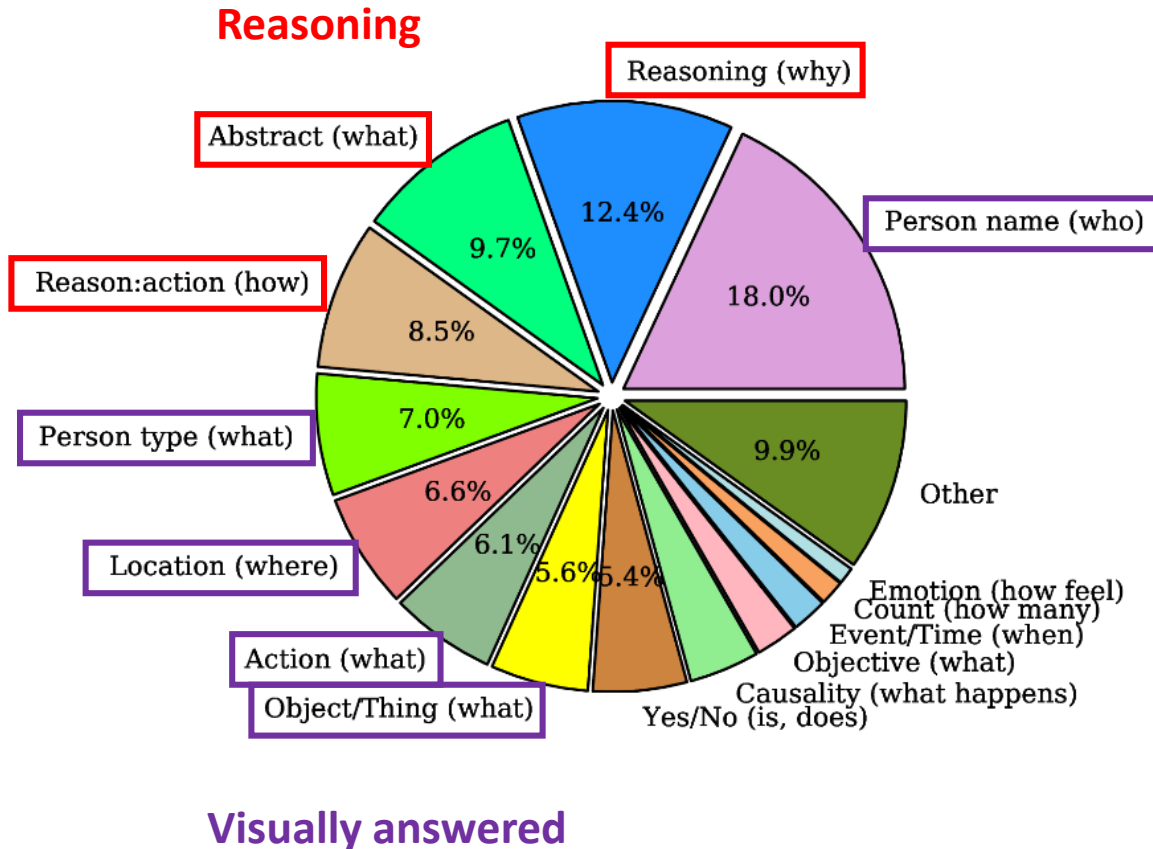
MovieQA: Dataset statistics

Plots and Subtitles

#Movies	408
#QA	14,944
Question #W	9.3 \pm 3.5
Correct ans #W	5.6 \pm 4.1
Wrong ans #W	5.1 \pm 3.9

Movies with Video Clips

#Movies	140
#QA	6,462
#Video clips	6,771
Clip dur (s)	202.7 \pm 216.2
Clip #shots	46.3 \pm 57.1

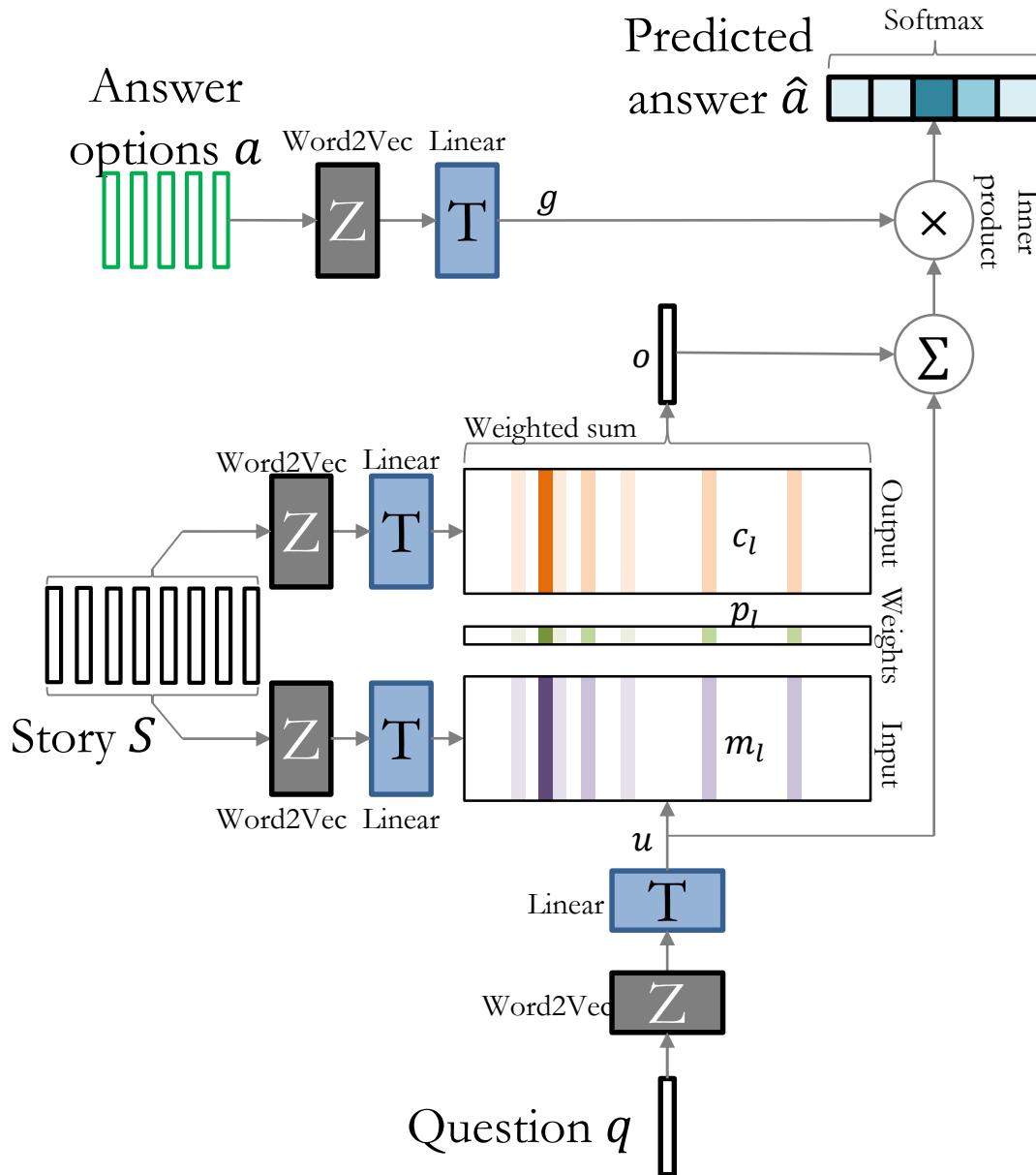


MovieQA:

Hasty answering, machine and human

- Don't look at the story, try to answer just based on question and answer
- Machine explores inherent dataset bias such as
 - Long answers are more likely to be correct
 - Most semantically similar answer is more likely to be correct
- Performs only slightly better (~25%) than random (20%)
- Human answering without story
 - Tests quality of confusing multiple choice options
 - Performance around 30%, indicates that answer options are quite likely, and it's not easy to pick the correct one

Tapaswi, et al. MovieQA: Understanding stories in movies through question-answering. In CVPR. 2016.



Memory Network

- Use fixed word embeddings Z (e.g. Word2Vec)
- Learn linear layers T to perturb these vectors
- Share weights T to reduce overfitting
- Allow for natural language answers
- Learn answer vectors g
- Scored with answer candidate o

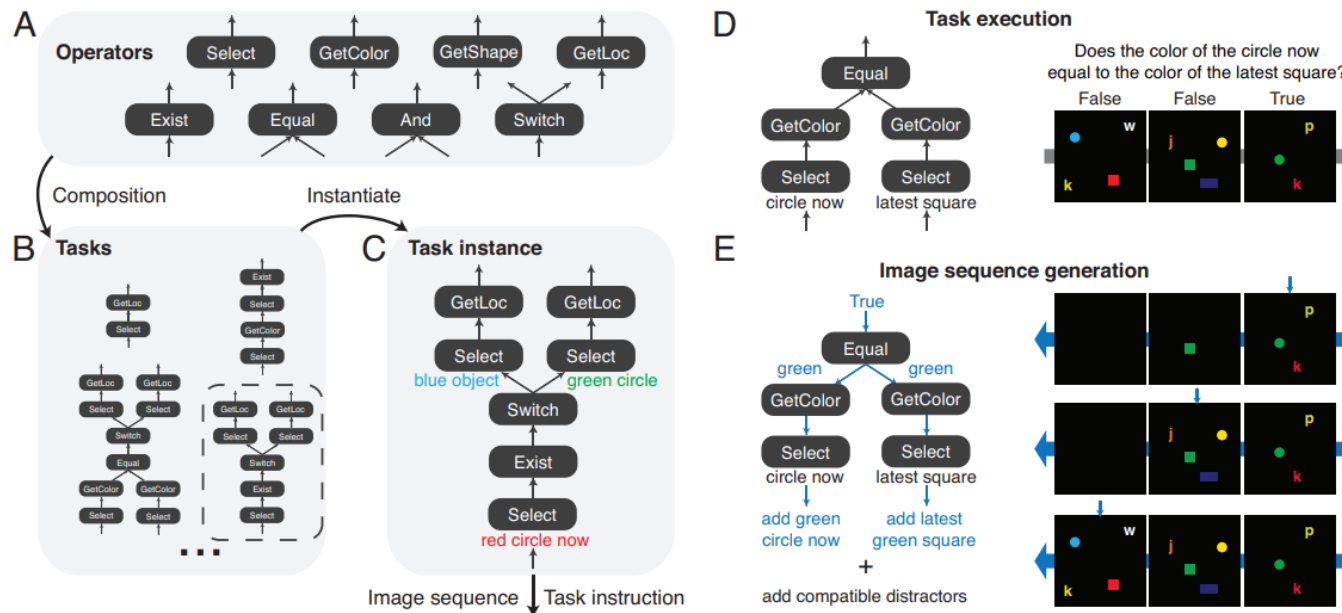
MovieQA evaluation

- Accuracy, fraction of correctly answered questions
- Random chance at 20% (5 multiple choice options)
- Answering without looking at the story is hard!
- Indicates small bias in dataset, truly deceiving multiple choice options
- Modified memory network shows decent performance
- Answering with all text sources is around 35-40% accuracy
- Answering with videos is very hard, performance near random (23%)
- A new benchmark challenge for teams around the world to try!

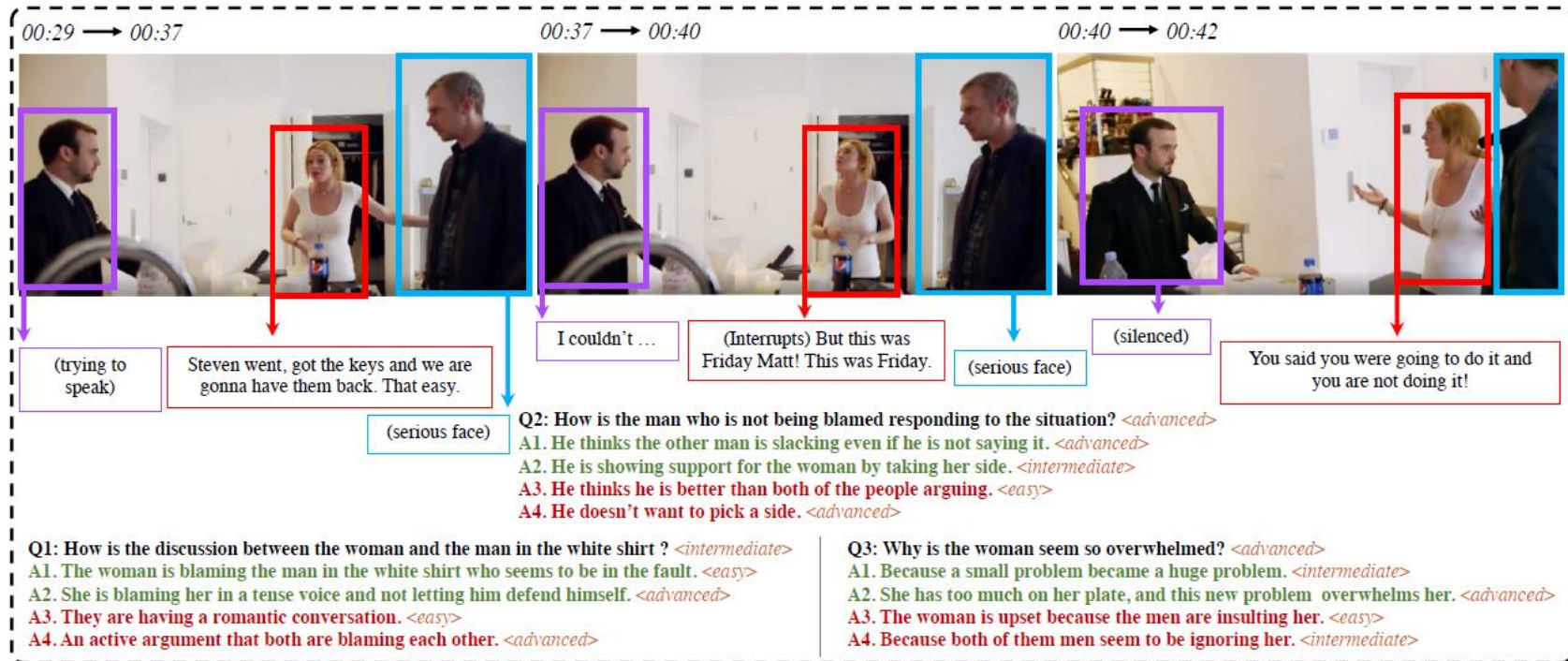
Tapaswi, et al. MovieQA: Understanding stories in movies through question-answering. In CVPR. 2016.

COG – Visual Cognition Dataset on Videos

- Video dataset using synthetic frames of letters and 2D shapes
- 44M questions about 11M videos
- Questions generated using several templates (see overview figure)



Yang, et al. A dataset and architecture for visual reasoning with a working memory. In ECCV. 2018.



- Social in-the-wild scenarios (from YouTube, carefully selected)
- 7.500 questions, 52.500 answers, different complexity levels
- Goal: train and evaluate socially intelligent systems

Summary

- QA – a good means to combine several levels of understanding
- Visual QA
 - Datasets: room scenes, fill in the blanks, VQA: image understanding
 - Different types of architectures: we presented four kinds
 - Neural Modules: Break the question into chunks, answer by reasoning
 - Spatial image attention to answer questions
 - Memory networks employ a memory and processing unit (reads, writes)
 - Graph networks use a structured representation of the image
- Video QA
 - Questions are complex
 - Requires high-level reasoning
 - Long term dependencies (many frames to answer some questions)
- A current hot topic in joint vision and language understanding
- Trend: Focus on *human interpretability* of the models

References – Datasets

■ Visual QA datasets

- DAQUAR: M. Malinowski and M. Fritz. A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input. In *NIPS* 2014
- Visual Madlibs: Yu, et al. Visual Madlibs: Fill in the Blank Image Generation and Question Answering. In *ICCV* 2015.
- VQA: S. Antol, et al. VQA: Visual Question Answering. In *ICCV* 2015.
- GQA: S. Antol, et al. VQA: Visual Question Answering. In *ICCV* 2015.
- CLEVR: S. Antol, et al. VQA: Visual Question Answering. In *ICCV* 2015.

■ VideoQA

- MovieQA: M. Tapaswi, et al. MovieQA: Understanding Stories in Movies through Question-Answering. In *CVPR* 2016.
- SocialIQ: A. Zadeh et al: Social-IQ: A Question Answering Benchmark for Artificial Social Intelligence. In *CVPR* 2019
- COG: A. Zadeh et al: Social-IQ: A Question Answering Benchmark for Artificial Social Intelligence. In *CVPR* 2019

References – VQA approaches

Global Embedding

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

Attention-based Techniques

- **SAN**: Yang et al. Stacked attention networks for image question answering. In CVPR 2016.
- **MCAN**: Z. Yu, et al. Deep Modular Co-Attention Networks for Visual Question Answering. In CVPR 2019.

Compositional Models

- **NMN**: J. Andreas, et al. Neural Module Networks. In CVPR 2016.
- **End-to-end NMN**: Hu, et al. Learning to reason: End-to-end module networks for visual question answering. In ICCV 2017.

Memory Networks

- **DMN**: Xiong, et al. Dynamic Memory Networks for Visual and Textual Question Answering. In ICML 2016.

Graph Nets

- **RN**: Santoro, et al. A simple neural network module for relational reasoning. In NIPS 2017.
- **GCN for FVQA**: Narasimhan, et al. Out of the box: Reasoning with graph convolution nets for factual visual question answering. In NIPS 2018.
- **Soft-Paths**: Haurilet et al. It's not about the Journey; It's about the Destination: Following Soft Paths under Question-Guidance for Visual Reasoning. In CVPR 2019.