Building neural network models that can reason

Stanford

Christopher Manning and Drew Hudson
Stanford University
@chrmanning ⋆ @stanfordnlp
But what about reasoning?
But what about reasoning?
But what about reasoning?
But what about reasoning?
But what about reasoning?
What is Reasoning? [Bottou 2011]
What is Reasoning? [Bottou 2011]

- Algebraically manipulating previously acquired knowledge in order to answer a new question
- Is not necessarily achieved by making logical inferences
- Continuity between algebraically rich inference and connecting together trainable learning systems
- Central to reasoning is composition rules to guide the combinations of modules to address new tasks
Worshipping the tabula rasa
Worshipping the tabula rasa

A good inductive bias improves your ability to learn (quickly and well)
Appropriate structural priors
Appropriate structural priors

- Convolution
- Attention
- Gating (LSTM/highway)
Tree-structured models

\[ p_2 = g(a, p_1) \]
\[ p_1 = g(b, c) \]

[Socher et al. 2010ff]
[Tai et al. 2015]
Tree-structured models

[Socher et al. 2010ff, Tai et al. 2015]
Tree-structured models

[Socher et al. 2011]
Compositional reasoning tree

The short-term memory (STM) holds a collection of saliency scores.
Compositional reasoning without trees
Compositional reasoning without trees
Compositional reasoning without trees

If $f: (X \times Y \times Z) \rightarrow N$, then $\text{curry}(f): X \rightarrow (Y \rightarrow (Z \rightarrow N))$
Our Goal

Rather than using standard machine learning correlation engines, the goal is improved neural network designs

- With a structural prior encouraging compositional and transparent multi-step reasoning
- While retaining end-to-end differentiability and demonstrated scalability to real-world problems
“When a person understands a story, [they] can demonstrate [their] understanding by answering questions about the story. Since questions can be devised to query any aspect of text comprehension, the ability to answer questions is the strongest possible demonstration of understanding.”

— Wendy Lehnert (PhD, 1977)
Talk Outline

- From Machine Learning to Machine Reasoning
  - MAC networks on the CLEVR task
    - The GQA dataset for VQA
    - Neural State Machines for VQA
There is a **purple cube** that is **behind** a **metal** object left to a **large ball**; what **material** is the cube? [Johnson et al, CVPR 2017]
There is a **purple cube** that is **behind** a **metal** object **left** to a **large ball**; what **material** is the cube?

[Johnson et al, CVPR 2017]
There is a purple cube that is behind a metal object left to a large ball; what material is the cube? Rubber [Johnson et al, CVPR 2017]
One Existing Approach...
Neural Module Networks

- Partially differentiable models that rely on strong supervision to translate queries into a tree-structured functional program
- The programs are used to compose a corresponding neural network out of a discrete collection of specialized neural modules

[Andreas at al, CVPR 2016; Johnson et al, ICCV 2017]
A neural model for problem solving and reasoning tasks

- Decomposes a problem into a sequence of explicit reasoning steps, each performed by a Memory-Attention-Composition (MAC) cell.
- One universal recurrent MAC cell is used throughout all the steps, where its behavior is versatile, adapting to the context in which it is applied.
- The network can represent arbitrarily complex reasoning graphs in a soft manner (self-attention), maintaining an end-to-end differentiability.
Memory, Attention, Composition. The MAC Network

Each **MAC cell** is responsible for performing **one reasoning step at a time**. It maintains **dual recurrent states**:

- **Control** $c_i$: this step’s **reasoning operation**
  
  *Attention-based average of a given query* (question)

- **Memory** $m_i$: **retrieved information** relevant to a query, accumulated over steps
  
  *Attention-based average of a given KB* (image)
Memory, Attention, Composition. The MAC cell.
Memory, Attention, Composition.

The MAC cell

- Control Unit (CU) computes a control state, extracting an instruction that focuses on some aspect of the query.
Memory, Attention, Composition.

The MAC cell

- **Control Unit (CU)** computes a **control** state, extracting an **instruction** that focuses on some aspect of the query.

- **Read Unit (RU)**: retrieves information from the **knowledge base** given the **current control state** and **previous memory**.
Memory, Attention, Composition. The MAC cell

- Control Unit (CU) computes a control state, extracting an instruction that focuses on some aspect of the query.
- Read Unit (RU): retrieves information from the knowledge base given the current control state and previous memory.
- Write Unit (WU): updates the memory state, merging old and new information.
The MAC cell
The Control Unit (CU)

Extract an instruction (control) from the question

control
$c_{i-1}$

query
$q$

contextual words
$cw_s$

Control Unit (CU)

1
$W, b$
$q_i$

2
$W, b$
$cq_i$

3
$W, b$

attention
softmax + weighted average

$c_i$
The MAC cell

The Read Unit (RU)

Retrieve information based on the current instruction

previous memory

$\mathbf{m}_{i-1}$

read unit (RU)

transitive reasoning

knowledge base

$\mathbf{KB}_{h,w}$

$m_i$ $\mathbf{W}, \mathbf{b}$

$\mathbf{W}, \mathbf{b}$

Compositional binary (and/or)

attention

$m_{new}$
The MAC cell

The Write Unit (WU)

Combine retrieved information with accumulated knowledge (memory)
The MAC net
From Cell to Network

A MacNet is a soft-attention sequence of $p$ MAC cells
The MAC net
From Cell to Network

A MacNet is a soft-attention sequence of $p$ MAC cells

Uniform sequential structure for all queries; efficient, easy to deploy, and fully differentiable
The MAC net
From Cell to Network

A MacNet is a soft-attention sequence of $p$ MAC cells

A capacity to represent arbitrarily complex reasoning Directed Acyclic Graphs (DAGs)
Experiments
CLEVR Overall Results

Overall Accuracy (0–100)

- 700k Training set
- 150k Test set
- 28 candidate answers
Experiments
CLEVR Overall Results

Overall Accuracy (0–100)

- 700k Training set
- 150k Test set
- 28 candidate answers

- Baseline: the most frequent answer for each question type

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE</td>
<td>41.8</td>
</tr>
<tr>
<td>CNN+LSTM</td>
<td>52.3</td>
</tr>
<tr>
<td>N2NMN</td>
<td></td>
</tr>
<tr>
<td>HUMAN</td>
<td></td>
</tr>
</tbody>
</table>
Experiments

CLEVR Overall Results

- 700k Training set
- 150k Test set
- 28 candidate answers

**Baseline**: the most frequent answer for each question type
Experiments
CLEVR Overall Results

Overall Accuracy (0–100)

- 700k Training set
- 150k Test set
- 28 candidate answers

Baseline: the most frequent answer for each question type
Experiments
CLEVR Overall Results

Overall Accuracy (95–100)

- (S): strongly supervised

<table>
<thead>
<tr>
<th></th>
<th>RN</th>
<th>PG+EE (S)</th>
<th>FILM</th>
<th>MACNET</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>99</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>98</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>97</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>96</td>
<td>95.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>95</td>
<td></td>
<td>96.9</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Experiments
CLEVR Overall Results

Overall Accuracy (95–100)

- (S): strongly supervised
- MAC net halves the previous best error rate
Existing Approaches
Relation Nets and FiLM

Large CNN stacks interleaved with specialized layers

RN [Santoro et al, 2017]
FiLM [Perez et al, 2017]
Existing Approaches

Relation Nets and FiLM

Large CNN stacks interleaved with specialized layers

- **Relation Net**: Inspects every pair of pixels in order to make predictions based on binary relations

RN [Santoro et al, 2017]
FiLM [Perez et al, 2017]
Existing Approaches
Relation Nets and FiLM

Large CNN stacks interleaved with specialized layers

- **Relation Net**: Inspects every pair of pixels in order to make predictions based on binary relations
- **FiLM**: Inserts conditional linear normalization layers that tilt the activations based on the question

RN [Santoro et al, 2017]
FiLM [Perez et al, 2017]
Experiments
CLEVR Overall Results

Overall Accuracy (95–100)

- (S): strongly supervised
Experiments
CLEVR Overall Results

Overall Accuracy (95–100)

- (S): strongly supervised
- MAC net halves the previous best error rate
Experiments
Data Efficiency

Learning curve

Accuracy (Val)

Data set size (% out of 700k train)

- MAC
- PG+EE (S)
- FiLM
- SA
Experiments
Data Efficiency

For 10% of the CLEVR dataset, 70k examples:
- **MacNet** achieves 86%
- Other approaches obtain 51.6% at best
- **Baseline** achieves 41.8%

Baseline
Most Frequent Answer for Question Type
# Experiments

## CLEVR-Humans

<table>
<thead>
<tr>
<th>CLEVR Humans</th>
<th>100</th>
<th>80</th>
<th>60</th>
<th>40</th>
<th>20</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN+LSTM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PG+EE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FILM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MACNET</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- CLEVR-Humans is **18k natural language questions** collected through **crowdsourcing**.
- They wrote "**questions hard for a smart robot to answer**".
- Dataset has **diverse vocabulary** and **linguistic variation**; demands more varied reasoning skills.
- Has a small training set for fine-tuning.
Experiments
CLEVR-Humans

CLEVR Humans

- CLEVR-Humans is 18k natural language questions collected through crowdsourcing
- They wrote "questions hard for a smart robot to answer"
- Dataset has diverse vocabulary and linguistic variation; demands more varied reasoning skills
- Has a small training set for fine-tuning
Experiments
CLEVR-Humans

CLEVR-Humans is 18k natural language questions collected through crowdsourcing.

They wrote “questions hard for a smart robot to answer”

Dataset has diverse vocabulary and linguistic variation; demands more varied reasoning skills.

Has a small training set for fine-tuning.
What color is the matte thing to the right of the sphere in front of the tiny blue block? **Purple**
Attention visualizations
A neural compositional reasoning engine

- An initial design for a compositional reasoning engine
  A constrained sequence model, separating control and memory and exploiting attention is a good prior for reasoning

- Strong compositional reasoning skills
  Halves the previous lowest error rate
  Generalizes much better from more modest training data
  Generalizes better to new tasks in CLEVR-Humans

- Generic, fully differentiable, end-to-end model
Earlier reasoning datasets are limited

Artificial images and/or language

A very small space of possible objects and attributes

High capacity models may memorize all combinations, reducing effective compositionality
Current VQA Benchmarks are problematic

Strong *language* real-world biases
models *guess* based on language priors

Visual biases
models overly focus on salient objects

Unclear error sources
*noisy* language; lack of object *grounding*

Little reasoning/compositionality required
GQA

a new dataset for compositional question answering over real-world images

- 10M compositional questions involving a diverse set of reasoning skills
- A balanced 1.5M-questions dataset with closely controlled answer distributions
GQA a new dataset for compositional question answering over real-world images

- Each **image** comes with a **scene graph** to represent its semantics
- Each **question** comes with a **functional program** to represent its semantics, grounded in the scene graph
GQA: a new dataset for compositional question answering over real-world images

- Questions are generated using a (traditional, rule-based) multi-step question engine focusing on linguistic diversity and a large vocabulary.
- A suite of new metrics exploit the known grounding to shed light on model behaviors in various aspects.
Visual Genome

[Krishna, Zhu, Groth, Johnson, Hata, Kravitz, Chen, Kalantidis, Li, Shamma, Bernstein, and Fei-Fei, IJCV 2017]
Visual Genome Scene Graph

[ Krishna, Zhu, Groth, Johnson, Hata, Kravitz, Chen, Kalantidis, Li, Shamma, Bernstein, and Fei-Fei, IJCV 2017 ]
Improved Visual Genome

- 108k images, each with a **Scene Graph** and object masks
- Use ontology of concepts: 1700 objects, 600 attributes and 330 relations, in 60 categories and subcategories
- Augment the graphs with (egocentric) **positional** (*left*), **comparative** (*same color*) and **global** information (*place*)
Question generation from graphs

Patterns: 500 probabilistic patterns, give a high-level question outline

What | Which <type> [do you think] <is> <dobject>, <attr> or <decoy>? 

Select: <dobject> → Choose <type>: <attr> | <decoy>
Example Questions

VQA
1. Does this man need a haircut?
2. What color is the guy’s tie?
3. What is different about the man’s suit that shows this is for a special occasion?

GQA
1. Is the person’s hair long and brown?
2. What appliance is to the left of the man?
3. Who is in front of the refrigerator on the left?
4. Is there a necktie in the picture that is not red?
5. Is the color of the vest different than shirt?
Baseline Accuracies

- Global Prior
- CNN
- Local Prior
- LSTM
- LSTM+ CNN
- Bottom Up
- MAC
- Human
Language VQA
Language

VQA

Language of Thought
Abstraction: Towards a Language of Thought

We see and reason with concepts, not visual details, 99% of the time. “Scene gists”

- A man
  - A cyclist
    - Wearing glasses, gloves, watch
- A cow
- Grassland
- Sky ... clouds
Abstraction: Towards a Language of Thought

- We use **concepts** to organize our sensory experience
- We build semantic **world models** relating concepts to represent our environment
- Used to **generalize** from given examples to new ones
- Used to draw **inferences** from facts to conclusions
The hope of deep neural models is to learn higher-level abstractions. Abstractions disentangle factors of variation, improving generalization.
Content-based attention over concepts

- Attention allows focus on a few elements out of a large set
- But we need attention over concept space, not over pixel space

- Cf. Yoshua Bengio’s so-called “Consciousness Prior”
  - Learn a deep representation that disentangles abstract explanatory factors
  - The conscious state is then a very low-dimensional vector, an attention mechanism applied on the deep representation
Learning by Abstraction: The Neural State Machine

[Hudson and Manning submitted]

- Operate over a vocabulary of **embedded concepts**, **atomic semantic units** that represent aspects of the world (the cleaned up Visual Genome ontology)
- **Translate** both **modalities** (image and question) to “speak the same language” of concepts
  - Everything is attention over the concept vocabulary
- **Abstract** over the raw dense features
- Inspired by **concept learning and use** in humans
A Neural State Machine

- A differentiable graph-based model that simulates the operation of a state machine
- Aims to combine the strengths of neural and symbolic approaches
A Neural State Machine

Two stages of construction and inference:

1) **Construction**: transforms the raw inputs into abstract semantic representations, building the state machine

   Image $\rightarrow$ Scene graph, Question $\rightarrow$ Instructions

2) **Inference**: simulates an iterative computation over the machine, sequentially traversing the states until completion.

   *Reasoning over the scene graph to compute an answer*
Formal Definition

- $C$ the model’s alphabet (embedded concepts)
- $S$ a set of states
- $E$ a set of edges for valid transitions
- $r_i$, $i \leq n$, instruction sequence
- $p_0$ distribution over the initial state
- $\delta: p_i \times r_i \rightarrow p_{i+1}$ a neural state transition function
Reasoning with Abstractions

Given an image, we construct a scene graph. Treat it as a neural state machine, where:
Reasoning with Abstractions

Given an image, we construct a scene graph. Treat it as a neural state machine, where:

- **States** correspond to **objects**
- **Transitions** correspond to **relations**
- States have different *(soft)* properties *(attributes)* via attention
Reasoning with Abstractions

Objects are represented through a factorized distribution over semantic properties (color, shape, material), defined over the concept vocabulary.
Reasoning with Abstractions

What is the red fruit inside of the bowl to the right of the coffee maker?

The question is translated into a series of instructions (with an attention-based encoder-decoder), defined over the concepts.
Reasoning with Abstractions

What is the red fruit inside of the bowl to the right of the coffee maker?

We simulate a computation as a neural state machine, feeding one instruction at a time and traversing the states until completion.
Reasoning with Abstractions

What is the red fruit inside of the bowl to the right of the coffee maker?

We simulate a computation as a neural state machine, feeding one instruction at a time and traversing the states until completion.
Reasoning with Abstractions

What is the red fruit inside of the bowl to the right of the coffee maker?

We simulate a computation as a neural state machine, feeding one instruction at a time and traversing the states until completion.
Reasoning with Abstractions

What is the red fruit inside of the bowl to the right of the coffee maker?

We simulate a computation as a neural state machine, feeding one instruction at a time and traversing the states until completion.
One more example

What is the **tall object** to the **left** of the **bed** made of?

- **Cabinet**: wood (0.95), tall (0.92), shiny (0.86)
- **Bed**: white (0.84), comfortable (0.91)
- **Lamp**: yellow (0.92), on (0.74), thin (0.82)

**Wood**
NSM accuracy on GQA

- CNN: 17.8
- Global Prior: 28.9
- Local Prior: 31.2
- LSTM: 41.1
- LSTM+ CNN: 46.6
- Bottom Up: 49.7
- MAC: 54.1
- NSM: 62.9

<table>
<thead>
<tr>
<th>Model</th>
<th>Dataset</th>
<th>Overall score</th>
</tr>
</thead>
<tbody>
<tr>
<td>d-LSTM Q + norm I</td>
<td>VQA v1</td>
<td>54.40</td>
</tr>
<tr>
<td>(Antol et al. ICCV 2015)</td>
<td>VQA-CP v1</td>
<td>23.51, -31%</td>
</tr>
<tr>
<td>NMN</td>
<td>VQA v1</td>
<td>54.83</td>
</tr>
<tr>
<td>(Andreas et al. CVPR 2016)</td>
<td>VQA-CP v1</td>
<td>29.64, -25%</td>
</tr>
<tr>
<td>SAN</td>
<td>VQA v1</td>
<td>55.86</td>
</tr>
<tr>
<td>(Yang et al. CVPR 2016)</td>
<td>VQA-CP v1</td>
<td>26.88, -29%</td>
</tr>
<tr>
<td>MCB</td>
<td>VQA v1</td>
<td>60.97</td>
</tr>
<tr>
<td>(Fukui et al. EMNLP 2016)</td>
<td>VQA-CP v1</td>
<td>34.39, -27%</td>
</tr>
</tbody>
</table>

What sport ...?
Generalization on VQA-CP v2

<table>
<thead>
<tr>
<th></th>
<th>SAN</th>
<th>HAN</th>
<th>GVQA</th>
<th>RAMEN</th>
<th>BAN</th>
<th>MuRel</th>
<th>ReGAT</th>
<th>NSM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank</td>
<td>24.96</td>
<td>28.65</td>
<td>31.30</td>
<td>39.21</td>
<td>39.31</td>
<td>39.54</td>
<td>40.42</td>
<td>45.80</td>
</tr>
</tbody>
</table>
# GQA Generalization Splits

<table>
<thead>
<tr>
<th>Structure</th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is the <code>&lt;obj&gt;</code> <strong>covered by</strong>?</td>
<td></td>
<td>What is covering the <code>&lt;obj&gt;</code>?</td>
</tr>
<tr>
<td>Is there a <code>&lt;obj&gt;</code> in the image?</td>
<td></td>
<td>Do you see any <code>&lt;obj&gt;</code>s in the photo?</td>
</tr>
<tr>
<td>What is the <code>&lt;obj&gt;</code> <strong>made of</strong>?</td>
<td></td>
<td>What material makes up the <code>&lt;obj&gt;</code>?</td>
</tr>
<tr>
<td><strong>What's the name</strong> of the <code>&lt;obj&gt;</code> <strong>that is</strong> <code>&lt;attr&gt;</code>?</td>
<td></td>
<td><strong>What is the</strong> <code>&lt;attr&gt;</code> <code>&lt;obj&gt;</code> <strong>called</strong>?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Content</th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Only questions that <strong>do not</strong> refer to any type of <strong>food</strong> or <strong>animal</strong> (do not have any word from these categories)</td>
<td></td>
<td>Only questions that refer to <strong>foods</strong> or <strong>animals</strong> (have a word from that one of these categories)</td>
</tr>
</tbody>
</table>
# GQA Generalization Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Content</th>
<th>Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global Prior</td>
<td>8.51</td>
<td>14.64</td>
</tr>
<tr>
<td>Local Prior</td>
<td>12.14</td>
<td>18.21</td>
</tr>
<tr>
<td>Vision</td>
<td>17.51</td>
<td>18.68</td>
</tr>
<tr>
<td>Language</td>
<td>21.14</td>
<td>32.88</td>
</tr>
<tr>
<td>Lang+Vision</td>
<td>24.95</td>
<td>36.51</td>
</tr>
<tr>
<td>BottomUp</td>
<td>29.72</td>
<td>41.83</td>
</tr>
<tr>
<td>MAC</td>
<td>31.12</td>
<td>47.27</td>
</tr>
<tr>
<td><strong>NSM</strong></td>
<td><strong>40.24</strong></td>
<td><strong>55.72</strong></td>
</tr>
</tbody>
</table>
VQA

Language of Thought
We should seek tasks involving understanding and multi-step compositional reasoning
Let’s build neural networks that think!

By iterative attention over abstracted, disentangled concepts
Tree-structured models

[Socher et al. 2011]
GQA a new dataset for compositional question answering over real-world images

- Questions are generated using a (traditional, rule-based) multi-step question engine focusing on linguistic diversity and a large vocabulary
- A suite of new metrics exploit the known grounding to shed light on model behaviors in various aspects
Visual Genome

[Krishna, Zhu, Groth, Johnson, Hata, Kravitz, Chen, Kalantidis, Li, Shamma, Bernstein, and Fei-Fei, IJCV 2017]
A Neural State Machine

Two stages of construction and inference:

1) **Construction**: transforms the raw inputs into **abstract** semantic representations, building the state machine

   Image $\rightarrow$ Scene graph, Question $\rightarrow$ Instructions

2) **Inference**: simulates an iterative computation over the machine, sequentially traversing the states until completion.

   *Reasoning over the scene graph to compute an answer*
Reasoning with Abstractions

Given an image, we construct a scene graph. Treat it as a neural state machine, where:

- **States** correspond to objects
- **Transitions** correspond to relations
- States have different *(soft)* properties *(attributes)* via attention