Learning By Abstraction:

The Neural State Machine

Drew Hudson & Christopher Manning

Stanford University

NeurIPS 2019
Language

At the conclusion he thrice prostrated: after which a solemn pause followed; after which Lama retired.

“One of the butler’s assistants other tied a scrap of red silk to saints), and the silk wrap; or

Chinlab (blessings connoisseurs)"
Language NLP

Information Extraction
Question Answering
Machine Translation
Summarization
Parsing
Dialogue
Language

Encodes information
Means of communication
Processing data, input
think through language, to represent information and experiences through a compositional discrete system
Thought itself posses a "language-like" compositional structure (*mentalese*), where simple concepts combine in systematic ways (akin to the rules of grammar in language) to build thoughts. In its most basic form, the theory states that thought, like language, has syntax.
Language of Thought

- **Generalization** (to a new concept, transfer)
- **Data Efficiency**: Learning from few examples
- **Interpretable**: Express our thought process
- **Abstraction**: support human-unique capability of reasoning and abstract thinking
Neural Networks

Continuous Computation

- They confuse **correlation** with causation
- They need a **lot of data** for training
- They **don’t generalize** to unseen conditions
- They are **hard to interpret**
Generalization without Systematicity: On the Compositional Skills of Sequence-to-Sequence Recurrent Networks

Brenden Lake\textsuperscript{1,2} Marco Baroni\textsuperscript{2}

Abstract

Humans can understand and produce new utterances effortlessly, thanks to their compositional skills. Once a person learns the meaning of a new verb “dax,” they can understand the meaning of “dax again.” This type of compositionality is central to the human ability to make strong generalizations from very limited data (Lake et al., 2017). In a set of influential and controversial papers, Jerry Fodor and other researchers have argued that the ability to make such strong generalizations must be non-compositional. We provide evidence that common neural network architectures are not actually able to make strong generalizations even given small amounts of training data. The problem is not with compositional generalization per se, but rather with the systematicity of generalization (i.e., the ability to make generalizations in a systematic manner).

\begin{verbatim}
jump => JUMP
jump left => LTURN JUMP
jump around right => RTURN JUMP
jump thrice => JUMP JUMP JUMP
jump opposite left and walk thrice => LTURN JUMP WALK WALK WALK
jump opposite left after walk around left => LTURN WALK LTTURN WALK LTTURN WALK LTTURN WALK LTTURN WALK
\end{verbatim}
Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet
Neural models are more biased towards texture and local patterns rather than shape and global structure.
Abstraction
Abstractions

• We form concepts to generalize from given examples to new ones
• build semantic world models to represent our environment
• draw inferences to proceed from facts to conclusions
The Neural State Machine

• A differentiable graph-based model the simulates the operation of a state machine.
• Combines the strengths of neural and symbolic approaches.
• Explore the model in the context of visual reasoning and question answering.
The Neural State Machine

Two stages of **learning** and **inference**:
1) **Modeling**: transforms the raw inputs into **abstract** semantic representations, and **construct the state machine**.

*Image → Scene graph, Question → 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

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

• The model operates over a vocabulary of **embedded concepts**, **atomic semantic units** that **represent** aspects of the world.

• **Translate** both modalities (image and question) to “**speak the same language**”.

• **Abstraction** over the raw dense features

• Inspired by **concept learning** in humans (cognitive science)
The Neural State Machine for VQA

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

- **States** correspond to objects
- **Transitions** correspond to relations.
- States have different *(soft)* properties *(attributes).*
Objects are represented through a **factorized distribution** over **semantic properties** (color, shape, material), defined over the **concept vocabulary**.
The question is translated into a series of instructions (with attention-based encoder-decoder), defined over the concepts.
Abstract Decoder

What color is the cat behind the red chair?

What **ATTR** is the **OBJ** **REL** the **ATTR** **OBJ**?

OBJ | ATTR | REL | OBJ | ATTR
---|---|---|---|---
chair | red | behind | cat | color

Disentangles structure and content
The Neural State Machine for VQA

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

We simulate a computation of the state machine, feeding one instruction at a time and traversing the states until completion.
Machine Simulation *(Traversal)*

\[
\gamma_i(s) = \sigma \left( \sum_{j=0}^{L} R_i(j)(r_i \circ W_j s^j) \right)
\]

\[
\gamma_i(e) = \sigma (r_i \circ W_{L+1} e')
\]

\[
p_{i+1}^s = \text{softmax}_{s \in S}(W_s \cdot \gamma_i(s))
\]

\[
p_{i+1}^r = \text{softmax}_{s \in S}(W_r \cdot \sum_{(s',s) \in E} p_i(s') \cdot \gamma_i((s', s)))
\]

\[
p_{i+1} = r_i' \cdot p_{i+1}^r + (1 - r_i') \cdot p_{i+1}^s
\]

Property type: Relation (0.92)

- \( r_i \)
- \( \gamma_e = 0.9 \)
- \( \gamma_s = 0.7 \)
- \( \gamma_s = 0.12 \)

- \( \gamma_e = 0.03 \)
- \( \gamma_e = 0.02 \)

Relation network:

- \( S \) to \( S' \) (holding)
- \( S \) to \( S' \) (behind)
- \( S \) to \( S' \) (looking)
- \( S \) to \( S' \) (holding)

Qualitative Results

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)

What is the **green food** inside of the **bowl**?
- **Apple**: yellow (0.58), round (0.75), healthy (0.81)
- **Broccoli**: green (0.94), leafy (0.93), fresh (0.92)
- **Bowl**: plastic (0.72), transparent (0.84)
Question Examples

1) What is the giraffe looking at? 
   person ✓
2) Is the fence in front of the giraffe made of metal? no ✓
3) Is the woman's shirt blue or yellow? blue ✓
4) On which side of the image is the person? right ✓
5) Is there a child behind the giraffe? no ✗

1) What is the fruit to the right of the salad? strawberries ✓
2) Is the fork to the right of the salad? no ✓
3) Is the plate white and square? no ✓
4) Is the cup behind the round plate? yes ✓
5) What is the plate made of? paper ✓

1) Are there either scarves or hats that are not pink? no ✓
2) Do the bear's dress and the person's shirt have the same color? yes ✓
3) Is the bear sitting or standing? sitting ✓
4) What is the green object that the bear is sitting on? book ✓
5) Is the bear wearing white shoes? yes ✓

1) Are there either a chair or a clock in the image? no ✓
2) Are there any flowers behind the bed on the left of the room? yes ✓
3) What color is the appliance on the right? black ✓
4) Is the carpet brown or blue? brown ✓
5) Is the TV turned on? yes ✓
Quantitative Results (GQA)
Generalization (VQA-CP)
### New Generalization Splits

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

Only questions that **do not** refer to any type of **food** or **animal** (do not have any word from these categories)

Only questions that refer to **foods** or **animals** (have a word from that one of these categories)
## 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>Lobal 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>
Summary

• **Construct** and **simulate** a neural state machine.

• A **neural traversal** over the **scene graph** guided by the **instructions** derived from the **questions** – **Sequential graph-based reasoning**.

• **Both** visual and linguistic **modalities** are transformed into the shared abstract language of concepts.

• Combines the strengths of **abstraction** and **compositionality**.

• A neural implementation of a classical model of computation!
Future Directions
Thank you! 😊