Neural Modular Networks

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

### Examples

- **Question:** How many different lights in various different shapes and sizes?
  **Answer:** Four

- **Question:** What is the color of the horse?
  **Answer:** Brown

- **Question:** What color is the vase?
  **Answer:** Green

- **Question:** Is the bus full of passengers?
  **Answer:** Yes

- **Question:** Is there a red shape above a circle?
  **Answer:** No

---

Figure 3: Example output from our approach on different visual QA tasks. The top row shows correct answers, while the bottom row shows mistakes (correct answers are given in parentheses).
Prior State of the Art

- **Semantic parsing and logic:**

![Diagram](image)

(a) Perception $f_{per}$ produces a logical knowledge base $\Gamma$ 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 $\ell$.

(c) Evaluation $f_{eval}$ evaluates a logical form $\ell$ on a logical knowledge base $\Gamma$ to produce a grounding $g$ and denotation $\gamma$.

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

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

1. Introduction
This paper describes an approach to visual question answering based on neural module networks (NMNs). We answer natural language questions about images using collections of jointly-trained neural “modules”, dynamically composed into deep networks based on linguistic structure. Concretely, given an image and an associated question (e.g. *where is the dog?*), we wish to predict a corresponding answer (e.g. *on the couch* or perhaps just *couch*). (Figure 1).

The visual QA task has significant applications to human-robot interaction, search, and accessibility, and has been the subject of a great deal of recent research [2, 7, 20, 22, 25, 32]. The task requires sophisticated understanding of both visual scenes and natural language. Recent successful approaches represent questions as bags of words, or encode the question using a recurrent neural network [22] and train a simple classifier on the encoded question and image. In contrast to these monolithic approaches, another line of work for textual QA [18] and image QA [21] uses semantic parsers to decompose questions into logical expressions. These logical expressions are evaluated against a purely logical representation of the world, which may be provided directly or extracted from an image [16].

In this paper we draw from both lines of research, presenting a technique for integrating the representational power of neural networks with the flexible compositional structure afforded by symbolic approaches to semantics. Rather than relying on a monolithic network structure to answer all questions, our approach assembles a network on the fly from a collection of specialized, jointly-learned modules (Figure 1). Rather than using logic to reason over truth values, we remain entirely in the domain of visual features and attentions.

Our approach first analyzes each question with a semantic parser, and uses this analysis to determine the basic composition of the question. The composition is then used to dynamically lay out a deep network composed of reusable modules. In visual question answering tasks, an additional sequence model provides sentence context and learns common-sense knowledge.

Use parser to extract logical expression from question.

Compose a neural network to answer question.

Reuse network components across problems.
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.

1. Introduction

This paper describes an approach to visual question answering based on neural module networks (NMNs). We answer natural language questions about images using collections of jointly-trained neural "modules", dynamically composed into deep networks based on linguistic structure. Concretely, given an image and an associated question (e.g. *where is the dog?*), we wish to predict a corresponding answer (e.g. *on the couch* or perhaps just *couch*) (Figure 1).

The visual QA task has significant applications to human-robot interaction, search, and accessibility, and has been the subject of a great deal of recent research [2,7,20,22,25,32]. The task requires sophisticated understanding of both visual scenes and natural language. Recent successful approaches represent questions as bags of words, or encode the question using a recurrent neural network [22] and train a simple classifier on the encoded question and image. In contrast to these monolithic approaches, another line of work for textual QA [18] and image QA [21] uses semantic parsers to decompose questions into logical expressions. These logical expressions are evaluated against a purely logical representation of the world, which may be provided directly or extracted from an image [16].

In this paper we draw from both lines of research, presenting a technique for integrating the representational power of neural networks with the flexible compositional structure afforded by symbolic approaches to semantics. Rather than relying on a monolithic network structure to answer all questions, our approach assembles a network on the fly from a collection of specialized, jointly-learned modules (Figure 1). Rather than using logic to reason over truth values, we remain entirely in the domain of visual features and attentions.

Our approach first analyzes each question with a semantic parser, and uses this analysis to determine the basic compositional structure of the question. This compositional structure is then used to dynamically compose a deep network from a collection of reusable modules. For visual question answering tasks, an additional sequence model provides sentence context and learns common-sense knowledge.

Question:

> Is there a circle next to a square?

Logical Expression:

> \textit{is}(circle, \texttt{next-to}(square))

- **Objective:** Convert question into logical expression.
- **Conceptually → Inducing a \textit{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
Visual question answering is fundamentally compositional, and the two main types of composition are: (1) combining "modules" into deep networks for question answering. Our approach uses a natural image dataset and a new dataset of complex questions to obtain structured representations of questions passed between modules.

Having built up an inventory of modules, we now need to assemble them into the layout specified by the question.

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

<table>
<thead>
<tr>
<th>Shapes Benchmark</th>
<th>size 4</th>
<th>size 5</th>
<th>size 6</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority</td>
<td>64.4</td>
<td>62.5</td>
<td>61.7</td>
<td>63.0</td>
</tr>
<tr>
<td>VIS+LSTM</td>
<td>71.9</td>
<td>62.5</td>
<td>61.7</td>
<td>65.3</td>
</tr>
<tr>
<td>NMN</td>
<td>89.7</td>
<td>92.4</td>
<td>85.2</td>
<td>90.6</td>
</tr>
<tr>
<td>NMN (easy)</td>
<td>97.7</td>
<td>91.1</td>
<td>89.7</td>
<td>90.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>VQA Benchmark</th>
<th>test-dev</th>
<th>test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes/No</td>
<td>Number</td>
</tr>
<tr>
<td>LSTM [2]</td>
<td>78.20</td>
<td>35.7</td>
</tr>
<tr>
<td>VIS+LSTM [2]</td>
<td>78.9</td>
<td>35.2</td>
</tr>
<tr>
<td>NMN</td>
<td>69.38</td>
<td>30.7</td>
</tr>
<tr>
<td>NMN+LSTM</td>
<td>77.7</td>
<td>37.2</td>
</tr>
<tr>
<td>Question</td>
<td>Network Structure</td>
<td>Correct Answer</td>
</tr>
<tr>
<td>-------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------</td>
<td>----------------</td>
</tr>
<tr>
<td>How many different lights in various different shapes and sizes?</td>
<td>measure(count(attend[light]))</td>
<td>four (four)</td>
</tr>
<tr>
<td>What is the color of the horse?</td>
<td>classify(color(attend[horse]))</td>
<td>brown (brown)</td>
</tr>
<tr>
<td>What color is the vase?</td>
<td>classify(color(attend[vase]))</td>
<td>green (green)</td>
</tr>
<tr>
<td>Is the bus full of passengers?</td>
<td>measure(is( combine(and(attend[bus], attend[full]))))</td>
<td>yes (yes)</td>
</tr>
<tr>
<td>Is there a red shape above a circle?</td>
<td>measure(is( combine(and(attend[red], re-attend<a href="attend%5Bcircle%5D">above</a>)) ))</td>
<td>no (no)</td>
</tr>
<tr>
<td>What is stuffed with toothbrushes wrapped in plastic?</td>
<td>classify(what(attend[stuff]))</td>
<td>container (cup)</td>
</tr>
<tr>
<td>Where does the tabby cat watch a horse eating hay?</td>
<td>classify(where(attend[watch]))</td>
<td>pen (barn)</td>
</tr>
<tr>
<td>What material are the boxes made of?</td>
<td>classify(material(attend[box]))</td>
<td>leather (cardboard)</td>
</tr>
<tr>
<td>Is this a clock?</td>
<td>measure(is(attend[clock]))</td>
<td>yes (no)</td>
</tr>
<tr>
<td>Is a red shape blue?</td>
<td>measure(is( combine(and(attend[red], attend[blue]))) )</td>
<td>yes (no)</td>
</tr>
</tbody>
</table>

Qualitative Results
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?

<table>
<thead>
<tr>
<th>Method</th>
<th>Overall</th>
<th>Exist</th>
<th>Count</th>
<th>equal</th>
<th>less</th>
<th>more</th>
<th>size</th>
<th>color</th>
<th>material</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN+BoW [26]</td>
<td>48.4</td>
<td>59.5</td>
<td>38.9</td>
<td>50</td>
<td>54</td>
<td>49</td>
<td>56</td>
<td>32</td>
<td>58</td>
</tr>
<tr>
<td>CNN+LSTM [4]</td>
<td>52.3</td>
<td>65.2</td>
<td>43.7</td>
<td>57</td>
<td>72</td>
<td>69</td>
<td>59</td>
<td>32</td>
<td>58</td>
</tr>
<tr>
<td>CNN+LSTM+MCB [9]</td>
<td>51.4</td>
<td>63.4</td>
<td>42.1</td>
<td>57</td>
<td>71</td>
<td>68</td>
<td>59</td>
<td>32</td>
<td>57</td>
</tr>
<tr>
<td>CNN+LSTM+SA [25]</td>
<td>68.5</td>
<td>71.1</td>
<td>52.2</td>
<td>60</td>
<td>82</td>
<td>74</td>
<td>87</td>
<td>81</td>
<td>88</td>
</tr>
<tr>
<td>NMN (expert layout) [3]</td>
<td>72.1</td>
<td>79.3</td>
<td>52.5</td>
<td>61.2</td>
<td>77.9</td>
<td>75.2</td>
<td>84.2</td>
<td>68.9</td>
<td>82.5</td>
</tr>
<tr>
<td>ours - policy search from scratch</td>
<td>69.0</td>
<td>72.7</td>
<td>55.1</td>
<td>71.6</td>
<td>85.1</td>
<td>79.0</td>
<td>88.1</td>
<td>74.0</td>
<td>86.9</td>
</tr>
<tr>
<td>ours - cloning expert</td>
<td>78.9</td>
<td>83.3</td>
<td>63.3</td>
<td>68.2</td>
<td>87.2</td>
<td>85.4</td>
<td>90.5</td>
<td>80.2</td>
<td>88.3</td>
</tr>
<tr>
<td>ours - policy search after cloning</td>
<td>83.7</td>
<td>85.7</td>
<td>68.5</td>
<td>73.8</td>
<td>89.7</td>
<td>87.7</td>
<td>93.1</td>
<td>84.8</td>
<td>91.3</td>
</tr>
</tbody>
</table>
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