Deep Compositional Question Answering With Neural Module Networks

GHIZAL SAHEBZAMANI, DELARAM BEHNAMI, PARDISS DANAEI, GOLARA JAVADI
19 MARCH 2019
Introduction

- Visual question answering based on neural module networks (NMNs)
- Understanding of both visual scenes and natural language

What color is the necktie? Yellow
Acknowledgement

- **PAPER 1**

- **PAPER 2**

- **PAPER 3**
Neural networks representation
- Represent question as bags of words
- Represent question using a recurrent neural network
- Train a simple classifier on the encoded question and image

Semantic Parsers

NMN: Assemble a network on the fly from a collection of jointly-learned modules. Analyze question with semantic parser. Determine computational units based on the analysis. Ground the...
NMN: Assemble a network on the fly from a collection of jointly-learned modules

- No single “best network” for all tasks
- Question answering is a highly-multitask learning setting
Ground the question on the image
Neural nets learn lexical groundings

Is there a red shape above a circle?

Yes

Use a recurrent network (LSTM) to read the question for final answer
Current Approach

- Ground the question on the image
  - Neural nets learn lexical groundings

Is there a red shape above a circle?

- Use a recurrent network (LSTM) to read the question for final answer

Yes
Neural Module Networks (NMNs)

- Model is a collection of modules
- Training datum
  - Natural language question
  - Image
  - Answer
- Classifier

Each module has a set of parameters, and a Network layout predictor.

Instantiate network based on network predictor.

**Is there a red shape above a circle?**

- red
- exists
- above

- true
Is there a red shape above a circle?

Yes.
Compositional VQA with NMNs

Breaking Down the Question to Its Semantic Building Blocks...

Off-the-shelf

V

VGG16

Where is the dog?

Parser

Layout

count

where

color

... 

dog

cat

standing

...

NMN

A

LSTM

couch

Single-layer (1024)
What Are the Building Blocks of Questions in VQA?

The 5 Modules

- classify
- combine
- attend
- re-attend
- measure
1. The **attend** Module for Attention to Predicates

- **Is there a red shape above a circle?**
- **Where is the dog?**

**attend**: \( \text{Image} \rightarrow \text{Attention} \)

- **attend[dog]**
- Convolution
2. The **re-attend** Module for Relationships

*Is there a red shape above a circle?*

**re-attend** : *Attention → Attention*

**re-attend[above]**

- FC
- ReLU ×2
3. The `combine` Module for Logical Operations

Is there a **red** circle?

**combine**: $Attention \times Attention \rightarrow Attention$

**combine[except]**

- Stack
- Conv.
- ReLU
4. The **measure** Module

Is there a red shape above a circle?

True

**measure**: $Attention \rightarrow Label$

measure[exists]

FC $\rightarrow$ ReLU $\rightarrow$ FC $\rightarrow$ Softmax $\rightarrow$ yes
5. The **classify** Module

What is the a red shape above a circle?

classify : $Image \times Attention \rightarrow Label$

classify[where]  

- Attend  
- FC  
- Softmax  
- couch
Training NMNs

Is there a red shape above a circle?

Diagram:

1. Be
2. Shape
3. Red
4. Above
5. Circle
6. A

Diagram:

1. Be
2. Shape
3. Red
4. Above
5. Circle
6. A

Diagram:

1. Exists
2. And
3. Shape
4. Red
5. Above
6. Circle
Training NMNs

- Learning parameters for individual module
- Supervised learning problem to maximize probability of output

\[ \arg \max_W \sum p(\text{yes} | \text{red}, \text{above}, \text{circle}; W) \]

Is there a red shape above a circle? Yes
What color is the shape to the right of a circle? Blue

exists and red above circle
shape right_of square
left_of circle

[Diagram showing relationships and logic]
VQA Dataset

What color is the necktie?  

SHAPES Dataset - Novel

Is there a red shape above a circle?  

Is a green shape above left of a red shape?  

[Antol et al. 2015]
Experiment Results - VQA

What color is she wearing?

<table>
<thead>
<tr>
<th>Model</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>
</tbody>
</table>

Ours

59.4
Experiment Results - SHAPES

- Size: number of modules needed to instantiate NMN
- NMN (easy): modified training set with no size-6 questions

<table>
<thead>
<tr>
<th>Method</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>

*Zhou

*Yang

Ours 90.6
<table>
<thead>
<tr>
<th>what color is the vase?</th>
<th>is the bus full of passengers?</th>
<th>is there a red shape above a circle?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>classify[color]</strong>(attend[vase])</td>
<td><strong>measure[is]</strong>( combine[and]( attend[bus], attend[full]) )</td>
<td><strong>measure[is]</strong>( combine[and]( attend[red], re-attend<a href="attend%5Bcircle%5D">above</a>))</td>
</tr>
<tr>
<td><strong>green</strong> (green)</td>
<td>yes (yes)</td>
<td>no (no)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>what material are the boxes made of?</th>
<th>is this a clock?</th>
<th>is a red shape blue?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>classify[material]</strong>( attend[box] )</td>
<td><strong>measure[is]</strong>( attend[clock] )</td>
<td><strong>measure[is]</strong>( combine[and]( attend[red], attend[blue]))</td>
</tr>
<tr>
<td><strong>leather</strong> (cardboard)</td>
<td>yes (no)</td>
<td>yes (no)</td>
</tr>
</tbody>
</table>
A Closer Look...

Pros:
- Combines advantages of:
  - Representation learning (like a neural net)
  - Compositionality (like a semantic parser)

Cons:
- Limited to produce arbitrary forms
A Closer Look…

Pros:
- Combines advantages of:
  - Representation learning (like a neural net)
  - Compositionality (like a semantic parser)

Cons:
- Limited to produce arbitrary forms
- Structure selection

Is there a red shape above a circle?
A Closer Look...

Pros:
- Combines advantages of:
  - Representation learning (like a neural net)
  - Compositionality (like a semantic parser)

Cons:
- Limited to produce arbitrary forms
- Structure selection
- Potential error in generated model

What beach city is there in Florida?

(wrong module behavior)
A Closer Look...

Pros:
- Combines advantages of:
  - Representation learning (like a neural net)
  - Compositionality (like a semantic parser)

Cons:
- Limited to produce arbitrary forms
- Structure selection
- Potential error in generated model
- End-2-End structure learning
Extension of the Work

- Learning to Reason: End-to-End Module Networks for Visual Question Answering
Extension of the Work

How many things are the same size as the ball?

Find
Ball

Relocate
Same size

Count
How many?

Four
Question RNN encoder

Decoder RNN with attention

How many things are the same size as the ball?

Four

Layout Policy
Find
Relocate
Count

Network builder

Module Network
Find
Ball
Relocate
Same size
Count
How many?
Extension of the Work

- Explainable Neural Computation via Stack Neural Module Networks
Extension of the Work

- Answer visual questions
- Ground referential expressions
Bibliography

Thank You

and Happy Spring