Transparency By Design
Closing the Gap Between Performance and Interpretability in Visual Reasoning

Presented by: Patrick Hayes
Why is Transparency Important?

- What happens when you can not explain why this happened?
- How do you convince lawmakers, judges, consumers that your model works?
- Transparency helps us debug and establish credibility.
Opaque Visual Question Answering

What is on the water?

Predicted top-5 answers with confidence:

- **boat** 86.134%
- **boats** 9.394%
- **people** 8.479%
- **flag** 8.378%
- **ship** 8.319%
Easy to Get Excited – Hard to Trust
How is this model making decisions?

Predicted top-5 answers with confidence:
- boat: 84.725%
- flag: 7.657%
- people: 6.008%
- ship: 0.604%
- surfboard: 0.530%
Does it always just say boat?

Predicted top-5 answers with confidence:
- yes: 77.523%
- no: 22.480%
- water: 0.000%
- unknown: 0.000%
- flag: 0.000%
How do Transparency by Design Networks Help?

- Take a complex task and break it down into simpler tasks.
- Give the user a way to visualize how well the network does on each sub task.
- Separates the simple logic from the complex logic.
How Does it Work?

- First parse the natural language question into a computational graph.
- The model that converts a natural language question into a computational graph is previous work from Johnson et al.
Dynamically Build a Neural Network

- Each node of the computation graph will represent one neural network module. A module is a small neural network used to perform a given logical operation.
- This means different questions will produce different neural network architectures.
- Each image in batch could be run through a very different network.
## What are the Different Neural Modules?

<table>
<thead>
<tr>
<th>Module Type</th>
<th>Operation</th>
<th>Language Analogue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attention</td>
<td>Attention × Stem → Attention</td>
<td>Which things are [property]?</td>
</tr>
<tr>
<td>Query</td>
<td>Attention × Stem → Encoding</td>
<td>What [property] is x?</td>
</tr>
<tr>
<td>Relate</td>
<td>Attention × Stem → Attention</td>
<td>Left of, right of, in front, behind</td>
</tr>
<tr>
<td>Same</td>
<td>Attention × Stem → Attention</td>
<td>Which things are the same [property] as x?</td>
</tr>
<tr>
<td>Comparison</td>
<td>Encoding × Encoding → Encoding</td>
<td>Are x and y the same [property]?</td>
</tr>
<tr>
<td>And</td>
<td>Attention × Attention → Attention</td>
<td>Left of x and right of y</td>
</tr>
<tr>
<td>Or</td>
<td>Attention × Attention → Attention</td>
<td>Left of x or right of y</td>
</tr>
</tbody>
</table>
What is an Attention?

- An attention is a mask that lets the network know which parts of the image are relevant for this operation.
- The dimensions of the attention are the same as the input image.
- The attention starts off as all 1s. This means the entire image is relevant. Many modules take an attention as input and output a new attention.
- The following images depict the output attentions of attending to objects with various qualities.
What is a Stem?

- A stem is set of features extracted from the input image
- The stem is passed as input to most modules
- Pre-trained state-of-the-art image classification networks are used to create the stem. This paper uses the first few blocks of resnet.
- Stem loses some spatial information
What is an Encoding?

- Encodings represent specific features of an object.
- For example, a query module takes a stem and an attention and outputs an encoding. You can think of these modules as image classifiers. The query model outputs either the color, material, or shape of the image.

<table>
<thead>
<tr>
<th>Color</th>
<th>Encoding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red</td>
<td>[0,0,1]</td>
</tr>
<tr>
<td>Blue</td>
<td>[0,1,0]</td>
</tr>
<tr>
<td>Green</td>
<td>[1,0,0]</td>
</tr>
</tbody>
</table>
# What are the Different Neural Modules?

<table>
<thead>
<tr>
<th>Module Type</th>
<th>Operation</th>
<th>Language Analogue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attention</td>
<td>Attention $\times$ Stem $\rightarrow$ Attention</td>
<td>Which things are [property]?</td>
</tr>
<tr>
<td>Query</td>
<td>Attention $\times$ Stem $\rightarrow$ Encoding</td>
<td>What [property] is $x$?</td>
</tr>
<tr>
<td>Relate</td>
<td>Attention $\times$ Stem $\rightarrow$ Attention</td>
<td>Left of, right of, in front, behind</td>
</tr>
<tr>
<td>Same</td>
<td>Attention $\times$ Stem $\rightarrow$ Attention</td>
<td>Which things are the same [property] as $x$?</td>
</tr>
<tr>
<td>Comparison</td>
<td>Encoding $\times$ Encoding $\rightarrow$ Encoding</td>
<td>Are $x$ and $y$ the same [property]?</td>
</tr>
<tr>
<td>And</td>
<td>Attention $\times$ Attention $\rightarrow$ Attention</td>
<td>Left of $x$ and right of $y$</td>
</tr>
<tr>
<td>Or</td>
<td>Attention $\times$ Attention $\rightarrow$ Attention</td>
<td>Left of $x$ or right of $y$</td>
</tr>
</tbody>
</table>
Some Modules are Simple

- The “And module” takes two attention masks does a simple element wise min operation.
- The “Or module” takes two attention masks does a simple element wise max operation.
Relate Module

- The relate module produces an attention based on some spatial feature. For example everything to the left of an object or everything behind an object.
- Requires global context but needs to maintain spatial information.
- Use dilated convolutions
Architecture Diagrams for Other Modules

Attention
- Attention
- Conv2D
- Conv2D
- Conv2D
- Element Wise Mult
- Feature Map
- Attention

Query
- Feature Map
- Conv2D
- Conv2D
- Element Wise Mult
- Feature Map
- Attention

Comparison
- Feature Map
- Conv2D
- Conv2D
- Concatenate
- Feature Map
- Feature Map
Examples Architecture for Specific Questions

Input Question: What color is the big metal object?

Input Question: How many objects are blue and metal?
How does Backpropagation Work?

- The output is a single word [“yes”, “no”, “blue”, “small”, “5”, ...]
- These single words are represented as one-hot-encodings
- The loss if the difference between the predicted output vector and the target output vector
- This loss is backpropagation through the network as it was defined for that input question. So if the question did not involve a “Relate module” then this training example will not update the parameters of the “Relate module”.
Qualitative Results

**Input Image**

**Intermediate Output Images**
- filter_color[cyan]
- filter_shape[cylinder]

**Input Question**
How many cyan cylinders are there?

**Output**
2
Qualitative Results

Input Image

Intermediate Output Images

Input Question
What color is the object behind the red cube and in front of the cyan cylinder?

Output
brown
Qualitative Results

**Input Image**
![Input Image](image)

**Input Question**
How many green spheres are there?

**Output**
1

**Intermediate Output Images**

- `filter_color[green]`
- `filter_shape[sphere]`
What is TbD + Reg?

- Original attention masks contained noise in background. This noise did not affect predictions but made interpreting the masks difficult. A small regularization term is used to get rid of the noise.
What is TbD + Reg + HRes?

- The stem is produced by passing the input image through the first few layers of a pre-trained image classifier. Each pooling layers means a loss in resolution. If stem is too coarse, then generate the stem with fewer pooling layers.
- An interpretable model is easier to debug!

An input image (left) and the attention mask produced by the model when asked to attend to the region behind the blue rubber object and in front of the large cyan rubber cylinder with $14 \times 14$ (middle) and $28 \times 28$ (right) input features.
Related Work

Hudson et al. contribute Compositional Attention Networks which visualize the attentions of the natural language process.
CLEVR Dataset

- Focus on specific types of questions.
- Remove biases from the dataset.
- No outside information needed.

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?
Q: How many objects are either small cylinders or red things?
# Experiments on CLEVR dataset

<table>
<thead>
<tr>
<th>Model</th>
<th>Overall</th>
<th>Count</th>
<th>Compare Numbers</th>
<th>Exist</th>
<th>Query Attribute</th>
<th>Compare Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMN [2]</td>
<td>72.1</td>
<td>52.5</td>
<td>72.7</td>
<td>79.3</td>
<td>79.0</td>
<td>78.0</td>
</tr>
<tr>
<td>N2NMN [12]</td>
<td>88.8</td>
<td>68.5</td>
<td>84.9</td>
<td>85.7</td>
<td>90.0</td>
<td>88.8</td>
</tr>
<tr>
<td>Human [17]</td>
<td>92.6</td>
<td>86.7</td>
<td>86.4</td>
<td>96.6</td>
<td>95.0</td>
<td>96.0</td>
</tr>
<tr>
<td>CNN + LSTM + RN [26]</td>
<td>95.5</td>
<td>90.1</td>
<td>93.6</td>
<td>97.8</td>
<td>97.1</td>
<td>97.9</td>
</tr>
<tr>
<td>PG + EE (700k) [18]</td>
<td>96.9</td>
<td>92.7</td>
<td>98.7</td>
<td>97.1</td>
<td>98.1</td>
<td>98.9</td>
</tr>
<tr>
<td>CNN + GRU + CBN [23]</td>
<td>97.6</td>
<td>94.5</td>
<td>93.8</td>
<td>99.2</td>
<td>99.2</td>
<td>99.0</td>
</tr>
<tr>
<td>DDRprog [29]</td>
<td>98.3</td>
<td>96.5</td>
<td>98.4</td>
<td>98.8</td>
<td>99.1</td>
<td>99.0</td>
</tr>
<tr>
<td>CAN [7]</td>
<td>98.9</td>
<td>97.2</td>
<td><strong>99.4</strong></td>
<td><strong>99.5</strong></td>
<td>99.3</td>
<td><strong>99.5</strong></td>
</tr>
<tr>
<td>TbD-net (Ours)</td>
<td>98.7</td>
<td>96.8</td>
<td>99.1</td>
<td>98.9</td>
<td>99.4</td>
<td>99.2</td>
</tr>
<tr>
<td>TbD + reg (Ours)</td>
<td>98.5</td>
<td>96.5</td>
<td>99.0</td>
<td>98.9</td>
<td>99.3</td>
<td>99.1</td>
</tr>
<tr>
<td>TbD + reg + hres (Ours)</td>
<td><strong>99.1</strong></td>
<td><strong>97.6</strong></td>
<td><strong>99.4</strong></td>
<td>99.2</td>
<td><strong>99.5</strong></td>
<td><strong>99.6</strong></td>
</tr>
</tbody>
</table>
Experiments on CLEVR-CoGenT

- CLEVR-CoGent is partitioned into two conditions. In Condition A all cubes are either gray, blue, brown. In Condition B all cubes are either red, green, or cyan.
- Tests how well a network disentangles shape from color.
- TbD networks regain high accuracies after fine-tuning.

<table>
<thead>
<tr>
<th></th>
<th>Train A</th>
<th>Train B</th>
<th>Fine-tune B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td>A</td>
</tr>
<tr>
<td>PG + EE [18]</td>
<td>96.6</td>
<td>73.7</td>
<td>76.1</td>
</tr>
<tr>
<td>TbD + reg (Ours)</td>
<td>98.8</td>
<td>75.4</td>
<td>96.9</td>
</tr>
</tbody>
</table>
Asymmetric Entangling

- Even though the only brown things the network has seen are cubes and the only cubes the network has seen are brown, gray, or blue. The network can identify brown cylinders as brown, but cannot identify green cubes as cubes.

<table>
<thead>
<tr>
<th>Predict Shape</th>
<th>Predict Color</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(\sqrt{A})$</td>
<td>$P(\sqrt{B})$</td>
</tr>
<tr>
<td>Train A</td>
<td>0.90</td>
</tr>
<tr>
<td>Fine-tune B</td>
<td>0.77</td>
</tr>
</tbody>
</table>

**attention[cube]**  
**attention[brown]**
Contributions of Transparency By Design

- Transparency By Design networks achieve state-of-the-art accuracies on CLEVR.
- Show that compositional visual attention provides powerful insight into model behavior.
- Propose a method to quantitatively evaluate the interpretability of visual attention mechanisms.
Limitations

- Limited set of operations.
- A network trained on a biased dataset is expected to have entangled feature representations, but it is still not easy to identify those entanglements.
- No visualizations for the Johnson et al Program Generator

Example of Limitation

<table>
<thead>
<tr>
<th>Input Image</th>
<th>Input Question</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image.png" alt="Image" /></td>
<td>How many objects are not gray?</td>
</tr>
</tbody>
</table>

Output

Error “not” is not in our vocabulary. Please use words in our vocabulary.
Future Work

- Applying Transparency By Design networks to more diverse datasets and questions. Including a not operation.
- Build a curious model. Show a model an image and have it generate questions that will best help it learn. The model could ask questions that disentangle its features.
- Build a model that could answer reasoning questions about short video clips.
- Frame the CLEVR dataset as a 3D world where the model could move around to acquire more information about occluded shapes.
Summary

- Visual Question Answering is an abstract research field. It is exciting, but models can easily trick people into believing the model is more advanced than it is.
- Transparency by Design networks achieve high accuracies and are easy to interpret and debug.
- They achieve transparency by visualizing the outputs of simple primitives with attention masks.