Visual Question Answering and Beyond

Aishwarya Agrawal
Ph.D. Candidate
School of Interactive Computing
VQA Task
VQA Task

What is the mustache made of?
What is the mustache made of?
What is the mustache made of?

AI System

bananas
Applications of VQA

• An aid to visually-impaired

   Is it safe to cross the street now?
Applications of VQA

- Surveillance
  What kind of car did the man in red shirt leave in?
Applications of VQA

• Interacting with personal assistants
  Is my laptop in my bedroom upstairs?
Outline

Overview of VQA
[ICCV’15, IJCV’16, AI Mag’16]

Problem with existing setup + models
[EMNLP’16]

Overcoming priors
• A new evaluation protocol [CVPR’18]
• A novel architecture [CVPR’18]
• A novel objective function [NIPS’18]

Beyond VQA
[Work in progress]
VQA Dataset

What color are her eyes?  
What is the mustache made of?

How many slices of pizza are there?  
Is this a vegetarian pizza?

Is this person expecting company?  
What is just under the tree?

Does it appear to be rainy?  
Does this person have 20/20 vision?
VQA Dataset

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?
VQA Dataset

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?
VQA Dataset

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

Fine-grained recognition
VQA Dataset

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?
VQA

- Multimodal inputs – Image and Question
- Details of the image
- Common sense + knowledge base
- Task-driven
- Holy-grail of automatic image understanding
VQA Dataset Stats

>0.25 million images

>0.76 million questions

~10 million answers
Please visit www.visualqa.org for more details.

Full release is out! 254,721 images, 764,163 questions, 9,934,119 answers. VQA challenge announced! The VQA evaluation server is now up.

What is VQA?

VQA is a new dataset containing open-ended questions about images. These questions require an understanding of vision, language and commonsense knowledge to answer.

- Over 250K images (MSCOCO and abstract scenes)
- 3 questions per image
- 10 ground truth answers per question
- 3 plausible (but likely incorrect) answers per question
- Open-ended and multiple-choice answering tasks
- Automatic evaluation metric

Subscribe to our group for updates!

Dataset

Details on downloading the latest dataset may be found on the download webpage.

October 2015: Full release (v1.0)

Real Images
- 204,721 MSCOCO images (all of current train/val/test)
- 614,163 questions
- 6,541,630 ground truth answers
- 1,842,489 plausible answers

Abstract Scenes
- 50,000 abstract scenes
- 150,000 questions
- 1,500,000 ground truth answers
- 450,000 plausible answers
- 250,000 captions

July 2015: Beta v0.9 release

June 2015: Beta v0.1 release

Contact: visualqa@gmail.com
Interest in VQA
(http://www.visualqa.org/)

13k page views/month during VQA Challenge 2018
Other VQA Datasets

- Visual Turing Test [Geman et al., PNAS 2014]
- DAQUAR [Malinowski & Fritz, NIPS 2014]
- COCO-QA [Ren et al., NIPS 2015]
- FM-IQA [Gao et al., NIPS 2015]
- Visual7W [Zhu et al., CVPR 2016]
- Visual Genome [Krishna et al., IJCV 2016]
- CLEVR [Johnson et al., CVPR 2017]
- VQA v2.0 [Goyal et al., CVPR 2017]
SOTA in VQA over the years

Accuracy on VQA v2

17%

Slide Credit: Yash Goyal
Outline

Overview of VQA
[ICCV’15, IJCV’16, AI Mag’16]

Problem with existing setup + models
[EMNLP’16]

Overcoming priors
• A new evaluation protocol [CVPR’18]
• A novel architecture [CVPR’18]
• A novel objective function [NIPS’18]

Beyond VQA
[Work in progress]
VQA models lack compositionality
Compositionality

Training

Q: What color is the plate?
A: Green

Q: What color are stop lights?
A: Red

Testing

Q: What color is the stop light?
A: Green

Q: What is the color of the plate?
A: Red
Q: What color are the safety cones?
Q: What color are the safety cones?

GT Ans: green

Predicted Ans: orange

Q: What color are the cones?

GT Ans: orange

Q: What color is the cone?

GT Ans: orange

Q: What color are the cones?

GT Ans: orange
VQA models lack compositionality

VQA models are driven by language priors in training data
Q: Are they military
A: yes

Q: Are they playing a game? A: yes

GT Ans: yes
VQA models lack compositionality

VQA models are driven by language priors in training data

VQA models lack sufficient image grounding
Looking at the Image

Q: What does the red sign say?

Predicted Ans: stop

Correct Response
Overview of VQA
[ICCV’15, IJCV’16, AI Mag’16]

Problem with existing setup + models
[EMNLP’16]

Overcoming priors
• A new evaluation protocol [CVPR’18]
• A novel architecture [CVPR’18]
• A novel objective function [NIPS’18]

Beyond VQA
[Work in progress]
Problem with existing setup + models

Q-type: What room is
Q: What room is this?
A: Kitchen

Train

Q-type: What room is

<table>
<thead>
<tr>
<th>Room</th>
<th>Q-type</th>
<th>Training Prior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kitchen</td>
<td>What room is</td>
<td>100</td>
</tr>
<tr>
<td>Living Room</td>
<td>What room is</td>
<td>0</td>
</tr>
<tr>
<td>Bedroom</td>
<td>What room is</td>
<td>0</td>
</tr>
<tr>
<td>Office</td>
<td>What room is</td>
<td>0</td>
</tr>
<tr>
<td>Dining Room</td>
<td>What room is</td>
<td>0</td>
</tr>
<tr>
<td>Other</td>
<td>What room is</td>
<td>0</td>
</tr>
</tbody>
</table>
Problem with existing setup + models

Train
Q: What room is this?
A: Kitchen

Test
Q: What room is this?
A: Bathroom

- **Q-type:** What room is

<table>
<thead>
<tr>
<th>Room</th>
<th>Training Prior</th>
</tr>
</thead>
<tbody>
<tr>
<td>kitchen</td>
<td>100</td>
</tr>
<tr>
<td>living room</td>
<td>20</td>
</tr>
<tr>
<td>bedroom</td>
<td>10</td>
</tr>
<tr>
<td>office</td>
<td>5</td>
</tr>
<tr>
<td>dining room</td>
<td>0</td>
</tr>
<tr>
<td>other</td>
<td>0</td>
</tr>
</tbody>
</table>

Diagram showing the distribution of room types for training and test sets.
Problem with existing setup + models

### Train

**Q:** What room is this?

**A:** Kitchen

### Test

**Q:** What room is this?

**A:** Bathroom

**Prediction:** Kitchen
Problem with existing setup + models

- IID splits $\rightarrow$ similar priors in train and test
- Memorization of priors does not hurt as much
- Problematic for benchmarking progress
Meet VQA-CP!

- New splits of the VQA v1 and VQA v2 datasets
- Visual Question Answering under Changing Priors (VQA-CP v1/v2)
VQA-CP Train Split

VQA-CP Test Split
Performance of VQA models on VQA-CP

<table>
<thead>
<tr>
<th>Model</th>
<th>Dataset</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>d-LSTM Q + norm I</td>
<td>VQA v1</td>
<td>54.40</td>
</tr>
<tr>
<td></td>
<td>VQA-CP v1</td>
<td>23.51</td>
</tr>
<tr>
<td>(Antol et al. ICCV15)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NMN</td>
<td>VQA v1</td>
<td>54.83</td>
</tr>
<tr>
<td></td>
<td>VQA-CP v1</td>
<td>29.64</td>
</tr>
<tr>
<td>(Andreas et al. CVPR16)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAN</td>
<td>VQA v1</td>
<td>55.86</td>
</tr>
<tr>
<td></td>
<td>VQA-CP v1</td>
<td>26.88</td>
</tr>
<tr>
<td>(Yang et al. CVPR16)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCB</td>
<td>VQA v1</td>
<td>60.97</td>
</tr>
<tr>
<td></td>
<td>VQA-CP v1</td>
<td>34.39</td>
</tr>
<tr>
<td>(Fukui et al. EMNLP16)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Outline

Overview of VQA
[ICCV’15, IJCV’16, AI Mag’16]

Problem with existing setup + models
[EMNLP’16]

Overcoming priors
• A new evaluation protocol [CVPR’18]
• A novel architecture [CVPR’18]
• A novel objective function [NIPS’18]

Beyond VQA
[Work in progress]
Grounded Visual Question Answering (GVQA) Model

- Inductive biases in model architecture to prevent relying on priors

- Designed to disentangle:
  - What can be said?

Q: What room is this?
Grounded Visual Question Answering (GVQA) Model

• Inductive biases in model architecture to prevent relying on priors

• Designed to disentangle:
  – What can be said?  
  
Q: What room is this?
Grounded Visual Question Answering (GVQA) Model

• Inductive biases in model architecture to prevent relying on priors

• Designed to disentangle:
  – What can be said?
  – What should be recognized?

Q: What room is this?
Grounded Visual Question Answering (GVQA) Model

- Inductive biases in model architecture to prevent relying on priors

- Designed to disentangle:
  - What can be said?
  - What should be recognized?

Q: What room is this?
GVQA

Visual Concept Classifier (VCC)

Concepts grouped into clusters

Answer Cluster Predictor (ACP)

LSTM

Answer Predictor (AP)

VQA Answers (998)

Q: What color is the dog?

black
GVQA

- Disentangles visual recognition from answer-type prediction
- Explicitly enforces visual grounding
- No direct pathway from question to final answer
### Results

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>Overall</th>
<th>Diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>VQA-CP v1</td>
<td>GVQA (Ours)</td>
<td>39.23</td>
<td>+12%</td>
</tr>
<tr>
<td></td>
<td>SAN (Yang et al. CVPR16)</td>
<td>26.88</td>
<td></td>
</tr>
<tr>
<td>VQA-CP v2</td>
<td>GVQA (Ours)</td>
<td>31.30</td>
<td>+6%</td>
</tr>
<tr>
<td></td>
<td>SAN (Yang et al. CVPR16)</td>
<td>24.96</td>
<td></td>
</tr>
</tbody>
</table>
Problem with existing setup + models

Train

Q: What room is this?
A: Kitchen

Test

Q: What room is this?
A: Bathroom

Prediction
Bathroom
Q: What color are the bananas?

Q-classifier: non yes/no
ACP: color
VCC: bananas green many food

Answer: green

GVQA’s output
GVQA’s output

Q: What is the most prominent ingredient?

Correct Ans: pasta
Outline

Overview of VQA
[ICCV’15, IJCV’16, AI Mag’16]

Problem with existing setup + models
[EMNLP’16]

Overcoming priors
• A new evaluation protocol [CVPR’18]
• A novel architecture [CVPR’18]
• A novel objective function [NIPS’18]

Beyond VQA
[Work in progress]
Overcoming Priors with Adversarial Regularization

- A simple drop-in regularizer
- Question embeddings should not encode the information about the exact answer
\[ \mathcal{L}_{VQA}(f, g, h) \approx -\frac{1}{N} \sum_{i=1}^{N} \log f(v_i, q_i)[a_i] \]
What color are the bananas?
A Base VQA Model

Visual Encoder

What color are the bananas?

Question Encoder

green

Answer Distribution

55
What color are the bananas?

Answer: yellow
What color are the bananas?

$\mathcal{L}_{QA}(f_Q, g) \approx -\frac{1}{N} \sum_{i=1}^{N} \log f_Q(q_i)[a_i]$
What color are the bananas?

\[ \mathcal{L}_{VQA}(f, g, h) \]

\[ \mathcal{L}_{QA}(f_Q, g) \]
\[
\min_{f,g,h} \max_{f_Q} \mathcal{L}_{VQA}(f,g,h) - \lambda_Q \mathcal{L}_{QA}(f_Q, g)
\]
\[
\min_{f,g,h} \max_{f_Q} \mathcal{L}_{VQA}(f, g, h) - \lambda_Q \mathcal{L}_{QA}(f_Q, g)
\]
$$\min_{f,g,h} \max_{f_Q} L_{VQA}(f,g,h) - \lambda_Q L_{QA}(f_Q,g) - \lambda_H L_H(f,g,h,f_Q)$$
$$\min_{f,g,h} \max_{f_Q} L_{VQA}(f, g, h) - \lambda_Q L_{QA}(f_Q, g) - \lambda_H L_H(f, g, h, f_Q)$$
<table>
<thead>
<tr>
<th>Model</th>
<th>VQA-CP v2 test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Overall  Yes/No Number Other</td>
</tr>
<tr>
<td><strong>SAN</strong></td>
<td><strong>24.96</strong>  <strong>38.35</strong> <strong>11.14</strong> <strong>21.74</strong></td>
</tr>
<tr>
<td>(Yang et al. CVPR16)</td>
<td></td>
</tr>
<tr>
<td>Model</td>
<td>VQA-CP v2 test</td>
</tr>
<tr>
<td>-----------------</td>
<td>---------------</td>
</tr>
<tr>
<td></td>
<td>Overall</td>
</tr>
<tr>
<td>SAN (Yang et al. CVPR16)</td>
<td>+2%</td>
</tr>
<tr>
<td>SAN + Q-Adv</td>
<td></td>
</tr>
<tr>
<td>Model</td>
<td>VQA-CP v2 test</td>
</tr>
<tr>
<td>------------------</td>
<td>------------------------------</td>
</tr>
<tr>
<td></td>
<td>Overall</td>
</tr>
<tr>
<td>SAN (Yang et al. CVPR16)</td>
<td>24.96</td>
</tr>
<tr>
<td><strong>Ours</strong> SAN + DoE</td>
<td>25.75</td>
</tr>
</tbody>
</table>

*SAN* + DoE shows an improvement of +1% compared to SAN.
<table>
<thead>
<tr>
<th>Model</th>
<th>VQA-CP v2 test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Overall</td>
</tr>
<tr>
<td>SAN (Yang et al. CVPR16)</td>
<td>24.96</td>
</tr>
<tr>
<td>SAN + Q-Adv</td>
<td>27.24</td>
</tr>
<tr>
<td>SAN + DoE</td>
<td>25.75</td>
</tr>
<tr>
<td>SAN + Q-Adv + DoE</td>
<td>33.29</td>
</tr>
</tbody>
</table>

+8%
<table>
<thead>
<tr>
<th>Model</th>
<th>VQA-CP v2 test Overall</th>
<th>VQA-CP v2 test Yes/No</th>
<th>VQA-CP v2 test Number</th>
<th>VQA-CP v2 test Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>GVQA</td>
<td>31.30</td>
<td>57.99</td>
<td>13.68</td>
<td>22.14</td>
</tr>
<tr>
<td>SAN (Yang et al. CVPR16)</td>
<td>24.96</td>
<td>38.35</td>
<td>11.14</td>
<td>21.74</td>
</tr>
<tr>
<td>SAN + Q-Adv</td>
<td>27.24</td>
<td>54.50</td>
<td>14.91</td>
<td>16.33</td>
</tr>
<tr>
<td>SAN + DoE</td>
<td>25.75</td>
<td>42.21</td>
<td>12.08</td>
<td>20.87</td>
</tr>
<tr>
<td>SAN + Q-Adv + DoE</td>
<td>33.29</td>
<td>56.65</td>
<td>15.22</td>
<td>26.02</td>
</tr>
<tr>
<td>Model</td>
<td>VQA-CP v2 test</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------------------</td>
<td>---------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Overall</td>
<td>Yes/No</td>
<td>Number</td>
<td>Other</td>
</tr>
<tr>
<td>GVQA</td>
<td>31.30</td>
<td>57.99</td>
<td>13.68</td>
<td>22.14</td>
</tr>
<tr>
<td>SAN (Yang et al. CVPR16)</td>
<td>24.96</td>
<td>38.35</td>
<td>11.14</td>
<td>21.74</td>
</tr>
<tr>
<td>SAN + Q-Adv</td>
<td>27.24</td>
<td>54.50</td>
<td>14.91</td>
<td>16.33</td>
</tr>
<tr>
<td>SAN + DoE</td>
<td>25.75</td>
<td>42.21</td>
<td>12.08</td>
<td>20.87</td>
</tr>
<tr>
<td>SAN + Q-Adv + DoE</td>
<td>33.29</td>
<td>56.65</td>
<td>15.22</td>
<td>26.02</td>
</tr>
<tr>
<td>UpDn (Anderson et al. CVPR18)</td>
<td>39.74</td>
<td>42.27</td>
<td>11.93</td>
<td>46.05</td>
</tr>
<tr>
<td>UpDn + Q-Adv</td>
<td>40.08</td>
<td>42.34</td>
<td>13.02</td>
<td>46.33</td>
</tr>
<tr>
<td>UpDn + DoE</td>
<td>40.43</td>
<td>42.62</td>
<td>12.19</td>
<td>47.03</td>
</tr>
<tr>
<td>UpDn + Q-Adv + DoE</td>
<td>41.17</td>
<td>65.49</td>
<td>15.48</td>
<td>35.48</td>
</tr>
</tbody>
</table>
Outline

Overview of VQA
[ICCV’15, IJCV’16, AI Mag’16]

Problem with existing setup + models
[EMNLP’16]

Overcoming priors
• A new evaluation protocol [CVPR’18]
• A novel architecture [CVPR’18]
• A novel objective function [NIPS’18]

Beyond VQA
[Work in progress]
Person: PICK UP A BIG RED BLOCK.
Computer: OK. (does it)
Person: GRASP THE PYRAMID.
Computer: I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.
Person: FIND A BLOCK WHICH IS TALLER THAN THE ONE YOU ARE HOLDING AND PUT IT INTO THE BOX.
Computer: BY "IT", I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING.
Computer: OK. (does it)
Person: WHAT DOES THE BOX CONTAIN?
Computer: THE BLUE PYRAMID AND THE BLUE BLOCK.
There is a yellow cube.
There is a yellow cube.

Task

Agent

- add object, cube, yellow, large, at (12,17)
- add object, cube, yellow, small, at (8,14)
- add object, cube, yellow, small, at (22,12)
Technical challenges of interest to us

- Reward Learning
- Rich Action Space
- Diverse Outputs
Domains

- MNIST Digit Painting
- 3D Scene Construction
Domain 1: MNIST Digit Painting (Task)

Draw 9.

Paint five.
Domain 1: MNIST Digit Painting (Dataset)

- Instructions paired with MNIST images (60K images)
- Instruction template -- <Action> <Class Label>
  - <Action> = “Draw” | “Put” | “Paint” | “Add” | “Create”
  - <Class Label> = numerical (“0”) / word form (“zero”)
Domain 1: MNIST Digit Painting (Environment and Action Space)

- Environment: *libmypaint* – painting library
Domain 1: MNIST Digit Painting
(Envionment and Action Space)

- Action Space:
  - end point of the brush (on 32 x 32 grid),
  - control point of the brush (on 32 x 32 grid),
  - pressure applied to the brush (10 levels),
  - brush size (4 levels),
  - binary flag -- draw stroke / skip
- Size of the action space -- 83,886,080
Domain 2: 3D Scene Construction (Task)

There is a green cylinder.

There is a large sphere.
Domain 2: 3D Scene Construction (Dataset)

- Instructions paired with 3D scene images (16,159 images)
- Instruction template: “There is a” <Attribute> <Shape>
  - Attribute: any color (8), any size (large, small)
  - Shape: any shape (sphere, cube, cylinder)
- Total possible unique instructions = (8+2)*(3) = 30
Domain 2: 3D Scene Construction (Environment and Action Space)

- Environment: 3D Editor
- Action Space:
  - location of the object (on 32 x 32 grid),
  - object shape (3 shapes),
  - object size (2 sizes),
  - object color (8 colors),
  - flag -- add object / modify object / skip
- Size of the action space -- 147,456
Overview of the approach

- Instruction
- Policy Network (Generator)
- Program
- Environment (Renderer)
- Intermediate Image
- Discriminator
- Final Image
- Reward
- Instruction
- Example Goal Image

Extending Ganin et al., ICML18
Overview of the approach

- **Reward learning**: discriminator is learning consistency between instruction and image
- **Diversity**: Action sampling from non-peaky distribution
Overview of the approach
Overview of the approach

Program

Policy Network (Generator)

Environment (Renderer)

Instruction

Empty Canvas

Program

Policy Network (Generator)

Instruction

Environment (Renderer)

Intermediate Image

Intermediate Image
Policy Network

Decoder

Last State

LSTM

CNN

CNN

CNN

MLP

Previous action

There is a red sphere.

Renderer

Decoder from Ganin et al., ICML18
Overview of the approach

Instruction → Policy Network (Generator) → Program → Environment (Renderer) → Final Image → Discriminator

Reward

Intermediate Image

Instruction

Example Goal Image
There is a red sphere.

Discriminator
Domain 1: MNIST Digit Painting

Create zero

Put 1

Paint two

Draw 3

Add four

Draw 5

Paint six

Put 7

Create eight

Add 9
Domain 1: MNIST Digit Painting

L2

Create zero

Put 1

Paint two

Draw 3

Add four

Draw 5

Paint six

Put 7

Create eight

Add 9

Discriminator

Create eight

Add 9

Draw five

Add four

Paint six

Put 1

Paint two

Create zero
Domain 2: 3D Scene Construction (Discriminator)

There is a small sphere.

There is a large cylinder.

There is a yellow cube.
Outline

Overview of VQA
[ICCV’15, IJCV’16, AI Mag’16]

Problem with existing setup + models
[EMNLP’16]

Overcoming priors
• A novel split [CVPR’18]
• A novel architecture [CVPR’18]
• A novel objective function [NIPS’18]

Beyond VQA
Thanks Rama!

A man holding a beer bottle with two hands and looking at it. A man in a white t-shirt looks at his beer bottle. A man with black curly hair is looking at a beer. A man holds a bottle of beer examining the label.

What color are her eyes?
What is the mustache made of?
Thanks!

Questions?