Incorporating Textual Resources to Improve Visual Question Answering

Jialin Wu
Thesis proposal
Committee members: Ray Mooney (advisor), Greg Durrett, David Harwath, Roozbeh Mottaghi and Dhruv Batra
Outline

• Introduction

• Background & Related Work
  • Problem Formulation
  • Base Systems
  • Related Tasks

• Completed Work
  • Generating Captions for VQA (ACL 2019)
  • Self-Critical Reasoning for Robust VQA (NeurIPS 2019)
  • Multi-Modal Answer Validation for Knowledge-based VQA (Under Review)

• Proposed Work
  • Short-term proposals
    • Breaking Down Visual Questions
    • Generating Answer Candidates
  • Long-term proposals
    • Verifying Textual Knowledge
    • Multi-modal Explanations for VQA
Outline

• Introduction
• Background & Related Work
  • Problem Formulation
  • Base Systems
  • Related Tasks
• Completed Work
  • Generating Captions for VQA (ACL 2019)
  • Self-Critical Reasoning for Robust VQA (NeurIPS 2019)
  • Multi-Modal Answer Validation for Knowledge-based VQA (Under Review)
• Proposed Work
  • Short-term proposals
    • Breaking Down Visual Questions
    • Generating Answer Candidates
  • Long-term proposals
    • Verifying Textual Knowledge
    • Multi-modal Explanations for VQA
Introduction

Problem description

• Given a natural language question and an image, the machine learning system needs to predict an answer to the question.

• Question: What color is the woman's jacket?

• Answer: red
**Introduction**

Different types of visual questions

- General visual questions
  - Question: Does this boy have a full wetsuit on?
  - Answer: Yes
  - Captions: A young man wearing a wetsuit surfing on a wave.
Introduction
Different types of visual questions

• Visual questions under changing Priors (VQA-CP)
  • Question: What utensil is pictured?
  • Answer: Fork
  • Explanations: There is a fork on the table.

Introduction
Different types of visual questions

- Commonsense visual questions
  - Question: Is this in an Asian country?
  - Answer: Yes
  - Explanations: Japanese words on the train and Japan is an Asian country.
Introduction
Different types of visual questions

• Knowledge-based visual questions

  • Question: Which movie featured a man in this position telling his life story to strangers?

  • Answer: Forrest Gump

  • Wikipedia: Forrest Gump narrated his life's story at the northern edge of Chippewa Square in Savannah, Georgia, as he sat at a bus stop bench.
Outline

• Introduction
• Background & Related Work
  • Problem Formulation
  • Base Systems
  • Related Tasks
• Completed Work
  • Generating Captions for VQA (ACL 2019)
  • Self-Critical Reasoning for Robust VQA (NeurIPS 2019)
  • Multi-Modal Answer Validation for Knowledge-based VQA (Under Review)
• Proposed Work
  • Short-term proposals
    • Breaking Down Visual Questions
    • Generating Answer Candidates
  • Long-term proposals
    • Verifying Textual Knowledge
    • Multi-modal Explanations for VQA
Related Work

VQA problem formulation

- Most systems frame VQA as a classification problem

Question: What color is the woman's jacket?
Outline

• Introduction
• Background & Related Work
  • Problem Formulation
  • **Base Systems**
  • Related Tasks
• Completed Work
  • Generating Captions for VQA (ACL 2019)
  • Self-Critical Reasoning for Robust VQA (NeurIPS 2019)
  • Multi-Modal Answer Validation for Knowledge-based VQA(Under Review)
• Proposed Work
  • Short-term proposals
    • Breaking Down Visual Questions
    • Generating Answer Candidates
  • Long-term proposals
    • Verifying Textual Knowledge
    • Multi-modal Explanations for VQA
Related Work
VQA base systems

• Bottom-Up and Top-Down Attention (Up-Down)

Question: What color is the woman's jacket?
Related Work
VQA base systems

• ViLBERT

What color is the woman's jacket?

Object detection (Bottom-Up)

Outline

• Introduction
• Background & Related Work
  • Problem Formulation
  • Base Systems
  • Related Tasks
• Completed Work
  • Generating Captions for VQA (ACL 2019)
  • Self-Critical Reasoning for Robust VQA (NeurIPS 2019)
  • Multi-Modal Answer Validation for Knowledge-based VQA (Under Review)
• Proposed Work
  • Short-term proposals
    • Breaking Down Visual Questions
    • Generating Answer Candidates
  • Long-term proposals
    • Verifying Textual Knowledge
    • Multi-modal Explanations for VQA
Related Work

Image captions

• Generating textual description of the image

There is a lady wearing a red jacket playing with two dogs

There is a dog playing with a frisbee next to a tennis ball
Related Work

Generating Visual Explanation

- Attention-based Visual Explanation

Question: What color is the woman's jacket?

Object detection (Bottom-Up) → Joint features → Attention (Top-Down)

- Red
- Green
- Blue

Peter Anderson, et al. Bottom-Up and Top-Down Attention for Image Captioning and VQA. In CVPR 2018
Related Work

Generating Visual Explanation

• Gradient-based Visual Explanation

Related Work
Generating Visual Explanation

• Gradient-based Visual Explanation

Related Work
Generating Visual Explanation

- Textual Explanation

Q: Is this a healthy meal?

A: No

...because it is a hot dog with a lot of toppings.

A: Yes

...because it contains a variety of vegetables on the table.

Related Work
Generating Visual Explanation

• Multimodal Explanation

Question: What sport is pictured? Answer: Surfing
Explanation: Because the man is riding a wave on a surfboard.

Outline

• Introduction
• Background & Related Work
  • Problem Formulation
  • Base Systems
  • Related Tasks
• Completed Work
  • Generating Captions for VQA (ACL 2019)
  • Self-Critical Reasoning for Robust VQA (NeurIPS 2019)
  • Multi-Modal Answer Validation for Knowledge-based VQA (Under Review)
• Proposed Work
  • Short-term proposals
    • Breaking Down Visual Questions
    • Generating Answer Candidates
  • Long-term proposals
    • Verifying Textual Knowledge
    • Multi-modal Explanations for VQA
Completed Work
Generate captions for VQA

- Less structural than visual features
- A young man is on his surf board with someone in the background.
  - Objects: [man, surf board, background]
  - Attributes: [man is young]
  - Relationships: [surfer on surf board, man with someone, someone in background]

A man on a blue surfboard on top of some rough water.
A young surfer in a wetsuit surfs a small wave. A young man rides a surf board on a small wave while a man swims in the background. A young man is on his surf board with someone in the background. A boy riding waves on his surf board in the ocean.
Completed Work
Generate captions for VQA

- Relevant captions are more likely to contain helpful knowledge for VQA

Human Captions:
1) A man on a blue surfboard on top of some rough water.
2) A young surfer in a wetsuit surfs a small wave.
3) A young man rides a surf board on a small wave while a man swims in the background.
4) A young man is on his surf board with someone in the background.
5) A boy riding waves on his surf board in the ocean.

Question 1: Does this boy have a full wetsuit on?
Caption: A young man wearing wetsuit surfing on a wave.

Question 2: What color is the board?
Caption: A young man riding a wave on a blue surfboard.
Completed Work
Generate captions for VQA

• Select most relevant captions for supervision

Does this boy have a full wetsuit on?

A young surfer in a wetsuit surfs a small wave

A man on a blue surfboard on top of some rough water

Yes
Completed Work
Generate captions for VQA

- Select most relevant captions for supervision

Does this boy have a full wetsuit on?

A young surfer in a wetsuit surfs a small wave

Yes

A man on a blue surfboard on top of some rough water
Completed Work
Generate captions for VQA

- Captions aided VQA system
Completed Work

Generate captions for VQA

• Use captions as features

Q: What is he doing?
Caption: A man is taking a picture of himself with a phone.
A: Taking picture.

Q: Is he wearing a hat?
Caption: A man with glasses and a hat on.
A: Yes.
Completed Work

Generate captions for VQA

• Use captions to adjust the Visual attention

Question: What colors is the surfboard?  Answer: yellow and red

Caption: A group of people standing next to yellow board.

Visual attention

Answer: Yellow and blue

Caption adjusted visual attention

Answer: Yellow and red
Completed Work
Generate image captions for VQA

• Results on VQA v2 Validation set
Outline

• Introduction
• Background & Related Work
  • Problem Formulation
  • Base Systems
  • Related Tasks
• Completed Work
  • Generating Captions for VQA (ACL 2019)
  • **Self-Critical Reasoning for Robust VQA (NeurIPS 2019)**
  • Multi-Modal Answer Validation for Knowledge-based VQA(Under Review)
• Proposed Work
  • Short-term proposals
    • Breaking Down Visual Questions
    • Generating Answer Candidates
  • Long-term proposals
    • Verifying Textual Knowledge
    • Multi-modal Explanations for VQA
Completed Work
Self critical reasoning for robust VQA

• Visual question under changing Priors

What utensil is pictured?

VQA System

Training Answer Distribution

knife fork

Knife

Fork
Completed Work
Self critical reasoning for robust VQA

• Force VQA to focus on what humans focus on
  • Construct a set of objects that humans focus on

There is a fork near the cake.
Human textual explanation

OR
Proposal object set $\mathcal{L}$

Human visual explanation
Completed Work
Self critical reasoning for robust VQA

• Force VQA to focus on what humans focus on

What utensil is pictured?

VQA System

Knife

Fork
Completed Work
Self critical reasoning for robust VQA

• Results

• VQA-CP dataset manually set the train and test set in very different distribution
Completed Work
Self critical reasoning for robust VQA

- Extract the most influential object for the right answer “Fork”
Completed Work
Self critical reasoning for robust VQA

- Force the object to contribute more to the correct answer.
Completed Work
Self critical reasoning for robust VQA

• Results
•

![VQA scores chart](image-url)
Outline

• Introduction
• Background & Related Work
  • Problem Formulation
  • Base Systems
  • Related Tasks
• Completed Work
  • Generating Captions for VQA (ACL 2019)
  • Self-Critical Reasoning for Robust VQA (NeurIPS 2019)
  • Multi-Modal Answer Validation for Knowledge-based VQA (Under Review)
• Proposed Work
  • Short-term proposals
    • Breaking Down Visual Questions
    • Generating Answer Candidates
  • Long-term proposals
    • Verifying Textual Knowledge
    • Multi-modal Explanations for VQA
Completed Work
Multi-modal Answer Validation

- Knowledge-based Visual Questions

Q: Which movie featured a man in this position telling his life story to strangers?
Baseline: Cloth
Ours: Forrest Gump

Q: Is this a healthy dish?
Baseline: No
Ours: Yes

Q: What breed of dog is the dog in this photo?
Baseline: Shepherd
Ours: Golden retriever
Completed Work
Multi-modal Answer Validation

• Validating answers instead of directly predicting them

What English city is famous for a tournament for the sport this man is playing?

The modern game of tennis originated in Birmingham, England, in the late 19th century as lawn tennis.

It is popular for sports fixtures and hosts several annual events including a free opera concert at the opening of the opera season, other open-air concerts, carnival and labour day celebrations, and the Copenhagen historic grand prix, a race for antique cars.

Wimbledon is notable for the longest running sponsorship in sports history due to its association with Slazenger who have supplied all tennis balls for the tournament since 1902.
Completed Work
Multi-modal Answer Validation

• Validating answers instead of directly predicting them

Top-K Answer Score
Completed Work
Multi-modal Answer Validation

- Knowledge retrieval

Q: Which movie featured a man in this position telling his life story to strangers?

\[ a_1 : \text{Forrest Gump} \]
\[ a_2 : \text{Speed} \]

\[ s_{qa_1} : \text{Forrest Gump featured a man in this position...} \]
\[ s_{qa_2} : \text{Speed featured a man in this position telling...} \]
Completed Work
Multi-modal Answer Validation

- Knowledge retrieval

Queries

- Movie
- Man
- Sitting Man
- :
- Strangers
- Forrest Gump
- Gump
- Speed
**Completed Work**

**Multi-modal Answer Validation**

- Knowledge retrieval

<table>
<thead>
<tr>
<th>Visual knowledge Pool</th>
<th>WikiPedia Pool</th>
<th>Concepts Pool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detected objects</td>
<td>A man is an adult male human.</td>
<td>A Gentleman is at a movie</td>
</tr>
<tr>
<td>Searched Images</td>
<td>The novel also features Gump as an astronaut, a professional wrestler, and a chess player.</td>
<td>Forrest Gump is a film</td>
</tr>
<tr>
<td></td>
<td>Forrest Gump narrated his life's story at the ... as he sat at a bus stop bench</td>
<td>Strangers is related to people</td>
</tr>
<tr>
<td></td>
<td>Speed is a 1994 American action thriller film directed by Jan de Bont ...</td>
<td></td>
</tr>
</tbody>
</table>
Completed Work
Multi-modal Answer Validation

• Knowledge Matching

WikiPedia Knowledge

<table>
<thead>
<tr>
<th>movie</th>
</tr>
</thead>
<tbody>
<tr>
<td>man</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>sitting man</td>
</tr>
</tbody>
</table>
Forrest Gump narrated his life’s ...
... sat at a bus stop bench.
Completed Work
Multi-modal Answer Validation

- Knowledge Matching

<table>
<thead>
<tr>
<th>ConceptNet Knowledge</th>
</tr>
</thead>
<tbody>
<tr>
<td>movie</td>
</tr>
<tr>
<td>sitting man</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>Forrest Gump</td>
</tr>
<tr>
<td>Forrest Gump is a film</td>
</tr>
</tbody>
</table>
Completed Work
Multi-modal Answer Validation

• Knowledge Matching

Visual knowledge
Completed Work
Multi-modal Answer Validation

• Answer Prediction and Validation
Completed Work
Multi-modal Answer Validation

• Results

• Performance on OKVQA dataset
Results

What is the complimentary color to the frisbee?

Because orange and blue are complementary colors, life rafts and life vests are traditionally orange, to provide the highest contrast and visibility when seen from ships or aircraft over the ocean.
Completed Work
Multi-modal Answer Validation

• Results

• Who is the official in this sport?

Umpire, related to, referee
Umpire, synonym, referee
Umpire, related to, baseball official

Because orange and blue are complementary colors, life rafts and life vests are traditionally orange, to provide the highest contrast and visibility when seen from ships or aircraft over the ocean.

In the Indian subcontinent, red is the traditional color of bridal dresses, and is frequently represented in the media as a symbolic color for married women.
Completed Work
Multi-modal Answer Validation

• Results

• What kind of lamp is this?
Outline

• Introduction
• Background & Related Work
  • Problem Formulation
  • Base Systems
  • Related Tasks
• Completed Work
  • Generating Captions for VQA (ACL 2019)
  • Self-Critical Reasoning for Robust VQA (NeurIPS 2019)
  • Multi-Modal Answer Validation for Knowledge-based VQA (Under Review)
• Proposed Work
  • Short-term proposals
    • Breaking Down Visual Questions
    • Generating Answer Candidates
  • Long-term proposals
    • Verifying Textual Knowledge
    • Multi-modal Explanations for VQA
Proposed Work

Breaking Down visual questions

- Questions with multiple aspects focusing on multiple modalities

General VQA
Q: What sport is being played??
A: Tennis

KB-VQA
Q: What other surfaces might this sport be played on?
A: Clay
Proposed Work

Breaking Down visual questions

- Questions with multiple aspects focusing on multiple modals

**General VQA:**
Q: What color is the bowl?
A: White

**KB-VQA:**
Q: the vegetable that garnishes this dish is nutritious for what human body part?
A: Eye
Proposed Work
Breaking Down visual questions

- Multiple knowledge sources need to interact with each other
  - Visual linking
    - The vegetable that garnishes this dish
  - Concepts:
    - Eyes are body parts
  - Facts:
    - Carrots are rich in beta-carotene, which the body utilizes to produce Vitamin A
    - Vitamin A has multiple functions: it is important for growth and development, for the maintenance of the immune system, and for good vision
Proposed Work
Breaking Down visual questions

• Breaking Down Visual Questions

Visual Linking

- Facts: Carrots are rich in beta-carotene...
- Concepts...

The vegetable

- Is good for

Garnishes

- Which body part

This dish

- Eye
- Leg
- Arm
Outline

• Introduction
• Background & Related Work
  • Problem Formulation
  • Base Systems
  • Related Tasks
• Completed Work
  • Generating Captions for VQA (ACL 2019)
  • Self-Critical Reasoning for Robust VQA (NeurIPS 2019)
  • Multi-Modal Answer Validation for Knowledge-based VQA (Under Review)
• Proposed Work
  • Short-term proposals
    • Breaking Down Visual Questions
    • Generating Answer Candidates
  • Long-term proposals
    • Verifying Textual Knowledge
    • Multi-modal Explanations for VQA
Proposed Work
Generating answer candidates

- Top predictions from baseline model is not perfect
- Ignoring ontology knowledge

What US Island is this activity most associated with?

Top-5 predictions: beach, surfboard, surf, california, usa.
Proposed Work
Generating answer candidates

• Top predictions from baseline model is not perfect
• Ignoring linguistic features

Is that a panda or a cat?

Top-5 predictions: domestic, domestic shorthair, calico, feline, bombay
Proposed Work

Generating answer candidates

• Top predictions from baseline model is not perfect

• Fail when the answer does not appear in the training data

• Facts: So in 1926 an American surfer named Tom Blake invented the very first, hollow surfboard.

When was this piece of sporting equipment invented?

Top-5 Predictions: 1940, 1806, 1902, 1945, 1790
Proposed Work

Generating answer candidates

- Converting VQA to textual QA for answer candidates generation.
  - Multimodal knowledge-based QA dataset size:
    - KB-VQA: 2,402 questions
    - FVQA: 5826 questions
    - OKVQA: 14055 questions
Proposed Work
Generating answer candidates

• Converting VQA to textual QA for answer candidates generation.
  • Textual knowledge-based QA dataset size:
    • Squad 2.0: more than 100k questions
    • HOTPOT-QA: more than 113k questions
    • Natural Questions: more than 300k questions
Proposed Work
Generating answer candidates

• Possible set of replacements
  • This vegetable
    • (Red) carrot
    • Cilantro
  • Garnishes this dish
    • None
Proposed Work
Generating answer candidates

- Learns to rank the combinations of replacements
  - The carrot that garnishes this dish is good for which body parts?
  - The carrot is good for which body parts?
  - The red carrot that garnishes this dish is good for which body parts?
  - The red carrot is good for which body parts?
  - The cilantro that garnishes the noodle is good for which body parts?
Outline

• Introduction
• Background & Related Work
  • Problem Formulation
  • Base Systems
  • Related Tasks
• Completed Work
  • Generating Captions for VQA (ACL 2019)
  • Self-Critical Reasoning for Robust VQA (NeurIPS 2019)
  • Multi-Modal Answer Validation for Knowledge-based VQA (Under Review)
• Proposed Work
  • Short-term proposals
    • Breaking Down Visual Questions
    • Generating Answer Candidates
  • Long-term proposals
    • Verifying Textual Knowledge
    • Multi-modal Explanations for VQA
Proposed Work
Verifying Textual Knowledge

• Image captioning suffers from object hallucination issues

TopDown: A kitchen with a stove and a sink.  
TopDown: A couple of cats laying on top of a bed.
Proposed Work
Verifying Textual Knowledge

- Retrieval and answer predictions are two separate steps

**Q:** How is this form of transportation powered?
**Q:** On what type of fuel source do these vehicles run on?

**Sample Knowledge for electricity:**
In parallel to the development of the bus was the invention of the electric trolleybus, typically fed through trolley poles by overhead wires

**Sample Knowledge for diesel:**
The most common power source for bus since the 1920s has been the diesel engine.
Proposed Work

Verifying Textual Knowledge

• Object checklist
  • Trolley poles ✅
  • Overhead wires ✅

Q: How is this form of transportation powered?

Sample Knowledge for electricity:
In parallel to the development of the bus was the invention of the electric trolleybus, typically fed through trolley poles by overhead wires.
Proposed Work

Verifying Textual Knowledge

- Object checklist
  - Stove ✓
  - Sink ×

TopDown: A kitchen with a stove and a sink.
Outline

• Introduction
• Background & Related Work
  • Problem Formulation
  • Base Systems
  • Related Tasks
• Completed Work
  • Generating Captions for VQA (ACL 2019)
  • Self-Critical Reasoning for Robust VQA (NeurIPS 2019)
  • Multi-Modal Answer Validation for Knowledge-based VQA (Under Review)
• Proposed Work
  • Short-term proposals
    • Breaking Down Visual Questions
    • Generating Answer Candidates
  • Long-term proposals
    • Verifying Textual Knowledge
    • Multi-modal Explanations for VQA
Proposed Work
Multi Explanations for VQA

• Textual explanation + object linking

Question: What sport is pictured?  Answer: Surfing
Explanation: Because the man is riding a wave on a surfboard.
Proposed Work
Multi Explanations for VQA

• Factual sentences + knowledge linking

• Facts:
  • So in 1926 an American surfer named Tom Blake invented the very first, hollow surfboard.

• Concepts:
  • hollow surfboard is a type of sporting equipment

When was this piece of sporting equipment invented?
Thanks for you attention!