

FALL 2020 CS 395T



Visual Question Answering

CS 395T: Topics in Natural Language Processing

October 1st, 2020

SHIVAM GARG AND JIEHAO LIAO

Department of Computer Science, The University of Texas at Austin

Why Language and Vision together?

- Language Grounding:
 - Language currently in DL models is just a mathematical construct.
 - Humans associate language symbols to real-life objects/concepts.
 - Connecting language symbols to image space, i.e. assigning a perceptual construct to language.



Man climbing mountain



Man climbing stairs

Language Grounding Tasks

- Today we will discuss:
 - Visual Question Answering and its variants
 - Visual Question Generation

Generating Visual Explanations

[Hendricks et al.](#), ECCV 2016

Generating natural language explanations for image classification systems.



This is a pine grosbeak because this bird has a red head and breast with a gray wing and white wing.



This is a Kentucky warbler because this is a yellow bird with a black cheek patch and a black crown.



This is a pied billed grebe because this is a brown bird with a long neck and a large beak.



This is an arctic tern because this is a white bird with a black head and orange feet.

Situation Recognition

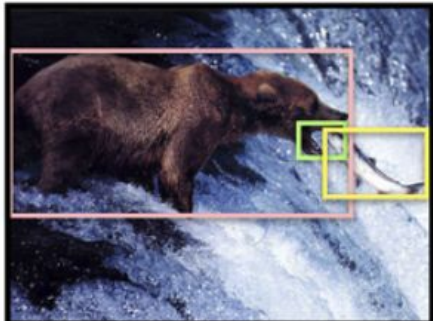
[Pratt et al.](#), ECCV 2016

Producing structured semantic summaries



Hitting

Agent	Tool	Victim	Victim Part	Place
Ballplayer	Bat	Baseball	∅	Field



Catching

Agent	Caught Item	Tool	Place
Bear	Fish	Mouth	River



Jumping

Agent	Source	Destination	Obstacle	Place
Female Child	Sofa	Sofa	∅	Living Room



Kneading

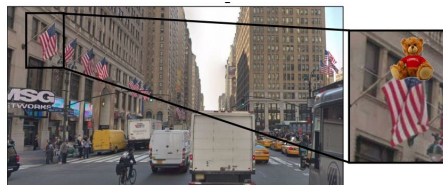
Agent	Item	Place
Person	Dough	Kitchen

Robotic Navigation

- Getting robots to have spatial reasoning in images from natural language such that they can navigate around the world and work on instructions.



Instruction: Head upstairs and walk past the piano through an archway directly in front. Turn right when the hallway ends at pictures and table. Wait by the moose antlers hanging on the wall.



Turn and go with the flow of traffic. At the first traffic light turn left. Go past the next two traffic light. As you come to the third traffic light you will see a white building on your left with many American flags on it. Touchdown is sitting in the stars of the first flag.

Goal: "Rinse off a mug and place it in the coffee maker"



Paper 1: Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, Devi Parikh

CVPR 2017

Visual Question Answering

- [Antol et al.](#), ICCV 2015
- QA task to answer the question given in natural language based on the image.
- Challenging:
 - Learn grounding b/w vision and language.
 - Open Domain already difficult in text based.
- 250k images(MS COCO + 50k abstract images)
- 750k questions, 10M answers



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?

Limitations of VQA

- Highly biased answer set.
 - “tennis” answer of “What is sport is” questions 41% of the time.
 - “2” answer of “How many” questions 39% of the time.
 - “yes” answer for “Do you see a ..” questions 87% of the time.
- Models tend to learn/focus on language aspect only by learning surface level word distribution.

Related Work

- Microsoft COCO-QA
 - Automatically generates QA pairs from the image captions.
 - 123k image-question pairs
 - 4 types of question templates: object, number, color, location
 - Answers are all one-word

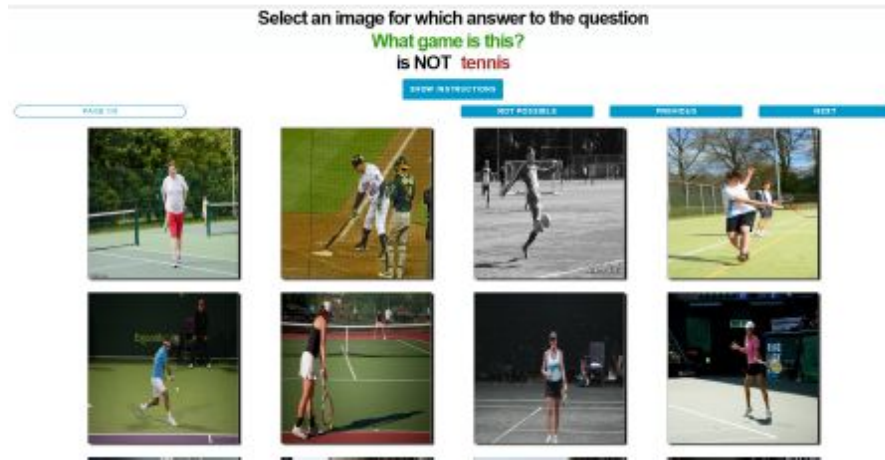
Source: <http://www.cs.toronto.edu/~mren/research/imageqa/data/cocoqa/>

Main Idea

- Augment VQA dataset so that image modality is needed to answer the question correctly.
- For each triplet (I, Q, A) in the dataset, introduce a triplet (I', Q, A') , s.t. I' is similar to I but the answer to question Q for the image I' is different from A .
- This would reduce the language bias in the dataset and the models will be forced to learn nuanced differences in the images to answer the questions correctly.

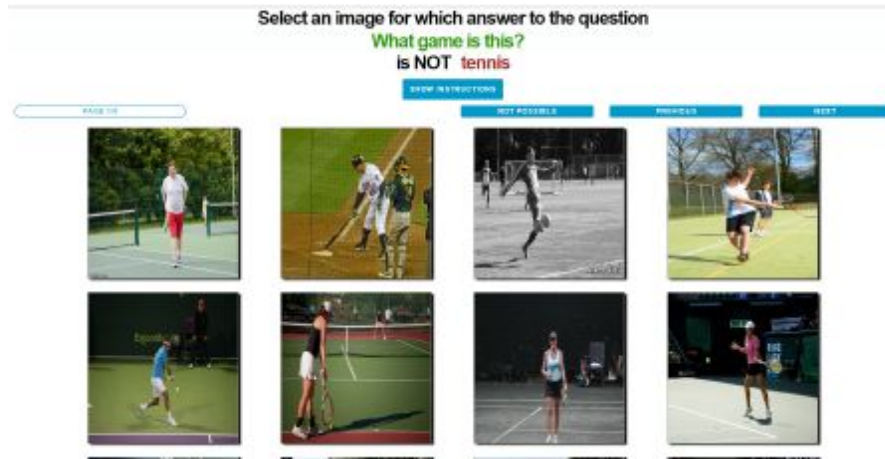
Annotation Process- Part I

- Dataset collected via Amazon Mechanical Turk.
- For each (I,Q,A) triplet, 24 nearest neighbours of I are selected.
- Turkers asked to select an image for which Q makes sense and answer is different from A.
- Nearest neighbours computed by taking the L2-Norm of the features extract from fc7 layer of [VGGNet](#).



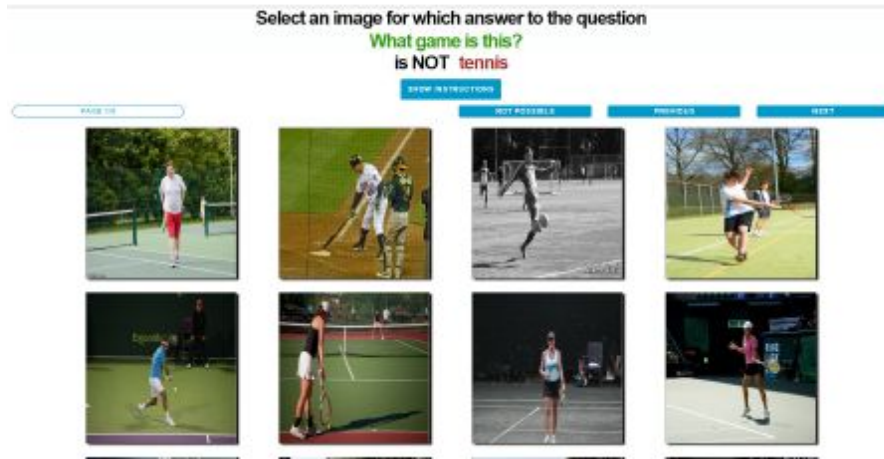
Annotation Process- Part I

- “Not Possible” option introduced.
 - The question does not make sense for all images in the candidate set
 - The answer among the candidate neighbours still remains A.
- Roughly 22% “Not Possible” selections.
 - Can be mitigated by introducing greater than 24 images, in scroll down behaviour.
- 193K complimentary images for train, 91K for val and 191K for test.



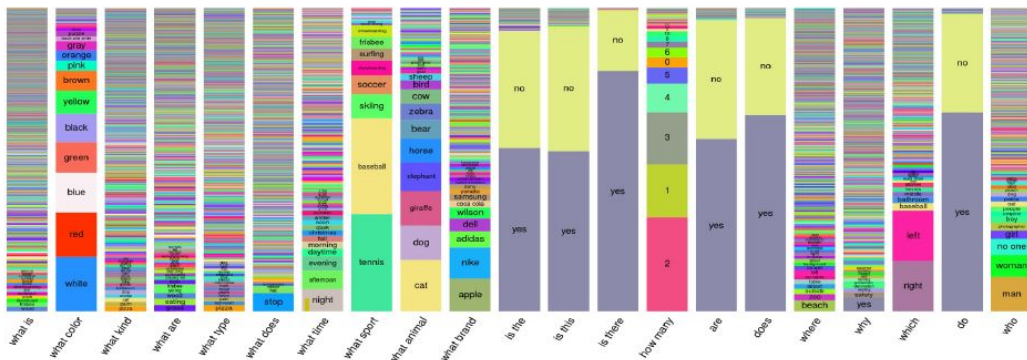
Annotation Process- Part II

- Complimentary images (I', Q) presented to 10 AMT workers
- The most common answer among the 10 answers is chosen as A'.
- 9% of the total annotations(A') still end up being same as A.
- Balanced VQA contains 443k train, 214k val and 453k test image-question pairs.

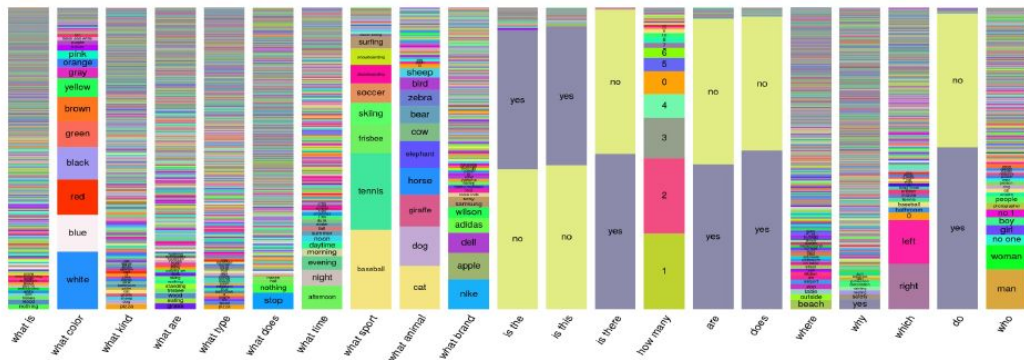


Answer Distribution

Answers from unbalanced dataset



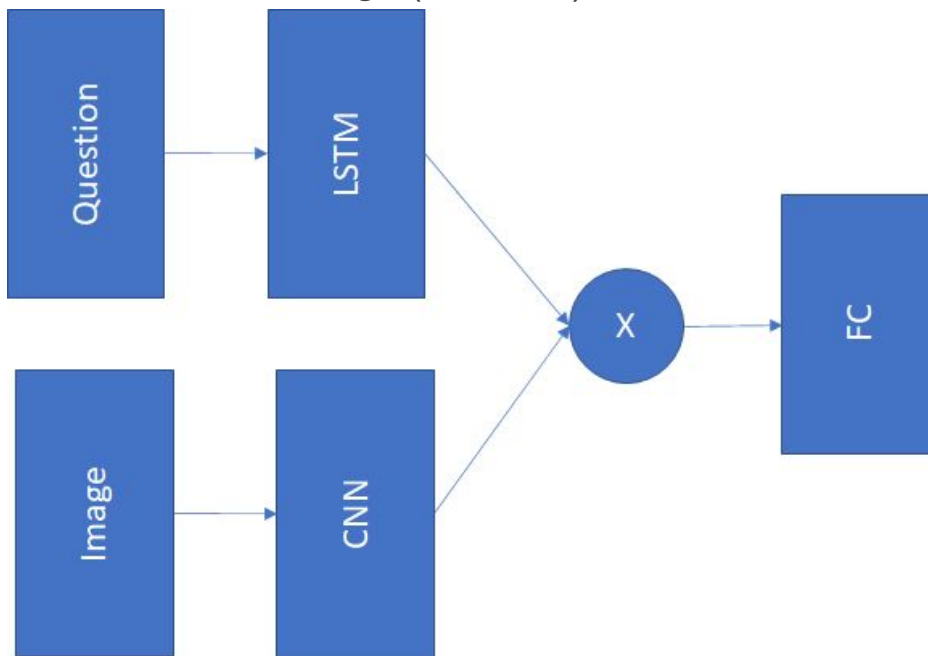
Answers from balanced dataset



- Certain categories balanced
 - Yes/No
 - Sports
 - Colours
 - Animals
 - Numbers
- Entropy of answer distribution increases by 56%.

VQA Models

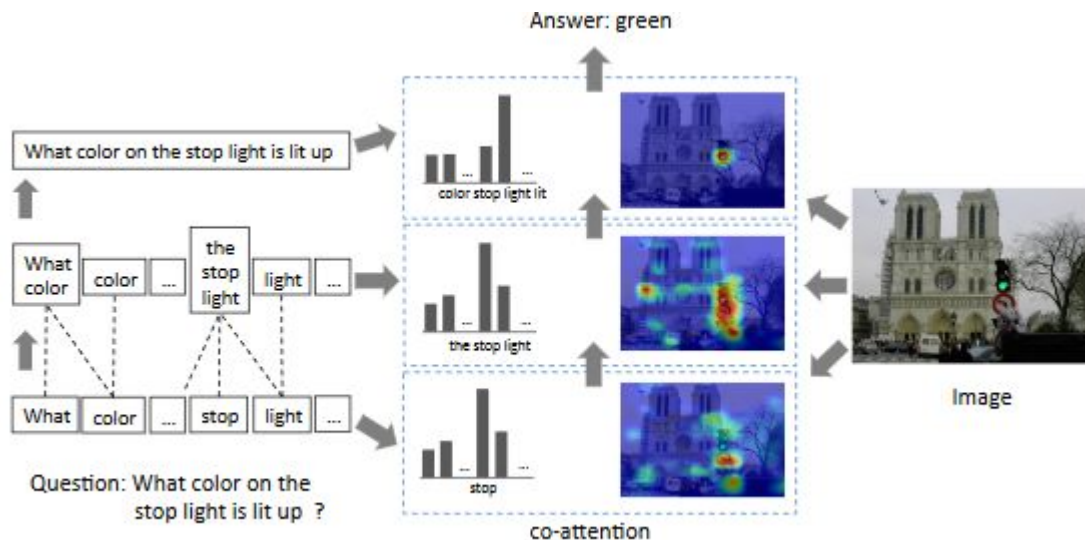
- Deep LSTM Question + norm Image ([Lu et. al.](#))



Answer chosen from 1000 most frequent answers in the training set.

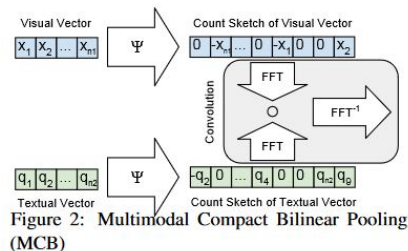
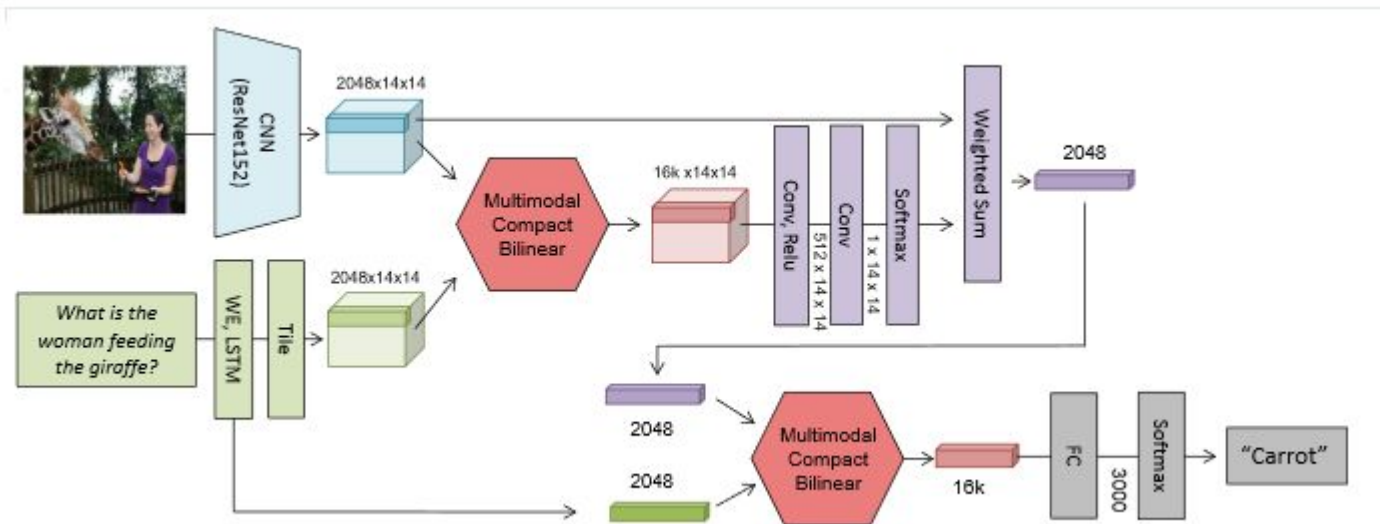
VQA Models

- Hierarchical Co-attention ([Lu et al.](#))
 - Attention based model in a hierarchical fashion at word-level, phrase-level and entire question-level.



VQA Models

- Multimodal Compact Bilinear Pooling ([Fukui et al.](#))
 - Multimodal compact bilinear pooling mechanism to attend over image features and combine them with language features.



VQA Models

- Two baseline models used for ablation study.
- Prior Model
 - Using the most frequent answer in the training set for each test question. The most common answer is “yes”.
- Language-only Model
 - Uses Deeper LSTM Question architecture without an image modality.

Experiments

- Uses the 5 VQA models to study the dataset properties.
 - Effect on performance when train set differs.
 - Benchmark performance on test-standard set
 - Performance analysis on different question types.

Experiments

- Models trained on VQA(original) underperform on the balanced test set.
 - Current models tend to exploit language biases.
 - Usage of balanced set leads to better performance, forcing the model to learn visual cues.
 - Using more training data increases test set performance, indicating extra data can help to improve the performance.
- Still a lot of scope for improvement.

Approach	UU	UB	B _{half} B	BB
Prior	27.38	24.04	24.04	24.04
Language-only	48.21	41.40	41.47	43.01
d-LSTM+n-I [24]	54.40	47.56	49.23	51.62
HieCoAtt [25]	57.09	50.31	51.88	54.57
MCB [9]	60.36	54.22	56.08	59.14

Table 1: Performance of VQA models when trained/tested on unbalanced/balanced VQA datasets. UB stands for training on Unbalanced train and testing on Balanced val datasets. UU, B_{half}B and BB are defined analogously.

Experiments

- Benchmark performance for existing models.
- Numerical category answers have the poorest performance.
 - Indicate the hardness of the model to count instances in the model.
- The absolute number are low, indicating a lot of scope for better modelling techniques.

Approach	All	Yes/No	Number	Other
Prior	25.98	61.20	00.36	01.17
Language-only	44.26	67.01	31.55	27.37
d-LSTM+n-I [24]	54.22	73.46	35.18	41.83
MCB [9]	62.27	78.82	38.28	53.36

Table 2: Performance of VQA models when trained on VQA v2.0 train+val and tested on VQA v2.0 test-standard dataset.

Experiments

- The performance of yes/no datasets drops significantly, indicating bias in the original training and validation set.
- $B_{\text{half}}B$ shows major improvement for yes/no and number answer types.
 - Indicates certain level of bias mitigation.

Approach	Ans Type	UU	UB	$B_{\text{half}}B$	BB
MCB [9]	Yes/No	81.20	70.40	74.89	77.37
	Number	34.80	31.61	34.69	36.66
	Other	51.19	47.90	47.43	51.23
	All	60.36	54.22	56.08	59.14
HieCoAtt [25]	Yes/No	79.99	67.62	70.93	71.80
	Number	34.83	32.12	34.07	36.53
	Other	45.55	41.96	42.11	46.25
	All	57.09	50.31	51.88	54.57

Table 3: Accuracy breakdown over answer types achieved by MCB [9] and HieCoAtt [25] models when trained/tested on unbalanced/balanced VQA datasets. UB stands for training on Unbalanced train and testing on Balanced val datasets. UU, $B_{\text{half}}B$ and BB are defined analogously.

Counter-example Explanation

- Proposing a model to predict the answer and output a counterexample explanation.
- Two step model proposed:
 - Given image I and question Q , predict answer A .
 - Given A and Q , find image I' , which is similar to I and has a different answer to Q .
- k -nearest neighbour images of I used to generate the candidate set of counter examples.
- The training data comes from the balanced VQA, since the data collection process collected similar images having different answers.



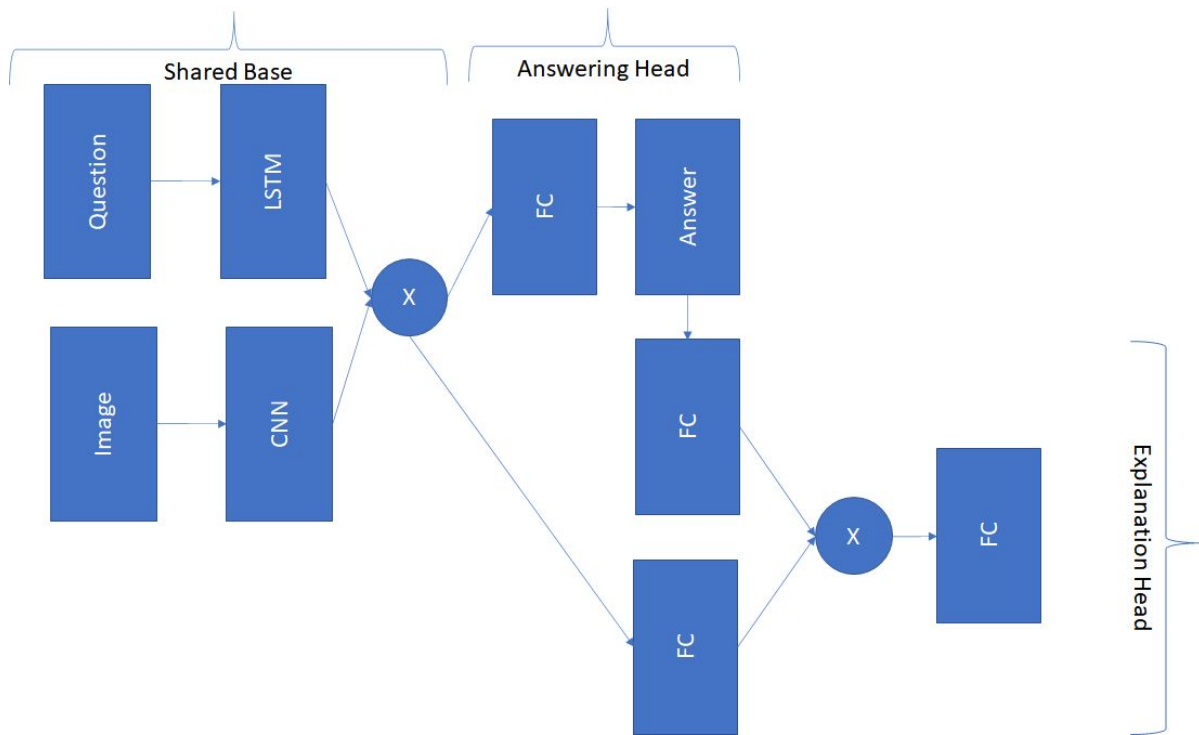
Q: Which way is its head turned?
A: left



Q: What color is the plate?
A: blue



Counter-example Explanation



Counter-example Explanation

- The explanation head is trained using contrastive loss(hinge loss). The net loss function of the whole network becomes:

$$\mathcal{L} = -\log P(A|I, Q) + \lambda \sum_i \max(0, M - (S(I') - S(I_i)))$$

- The positive and negative images for counter example come from the design of balanced VQA, where human selected images and random images are positive and negative examples respectively.

Counter-example Explanation

- Baselines:
 - Random
 - Distance: Nearest image in L2-norm space.
 - VQA Model: Image with least $P(A|Q,I')$ for a given VQA model.

	Random	Distance	VQA [3]	Ours
Recall@5	20.79	42.84	21.65	43.39

Contributions

- Bias mitigation of VQA dataset by collecting complementary similar images for each image-question pair which has a different answer.
- An annotation interface for hard negative example mining for data augmentation.
- Ablation study on various approaches for VQA on the new dataset .
- An interpretable QA system which uses counterexamples as the justification for the chosen answer.

Related Work

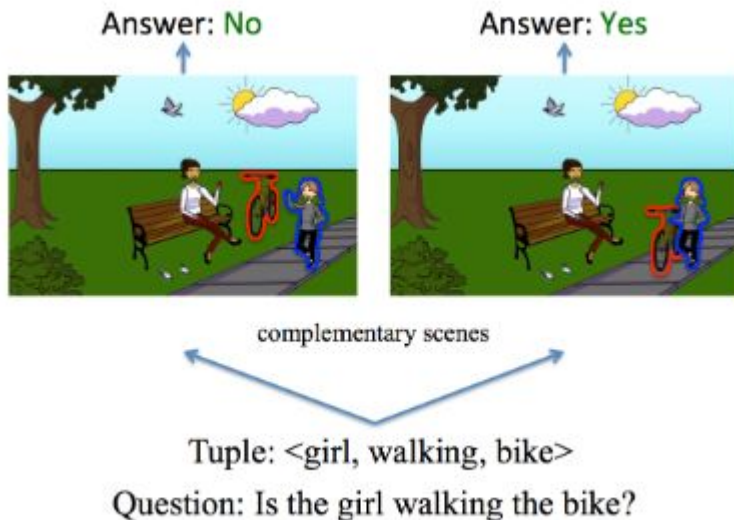
[Hodosh et al.](#), ACL 2016 implemented hand-designed rules to create two similar captions for images.

"Switch People" Task		
Image	Gold Caption	Distractor
	a man holding and kissing a crying little boy on the cheek	a crying little boy holding and kissing a man on the cheek
	a woman is hula hooping in front of an audience	an audience is hula hooping in front of a woman

Figure 1: The "switch people" task

Related Work

[Zhang et al.](#), CVPR 2016 ask human annotators to modify clipart images so that the answer to the question changes.



Related Work

[Goyal et al.](#), ICML 2016 explain VQA systems by giving visual explanations or spatial maps overlaid on images.



Question : What **vegetable** is on the plate ?

Predicted Answer : broccoli

Question : What **color** is the plate ?

Predicted Answer : white

Question : Is there meat in this dish ?

Predicted Answer : no



Question : **Where** is the player ?

Predicted Answer : tennis court

Question : **what** does the man wear on his arms ?

Predicted Answer : tennis racket

Question : **what** sport is this ?

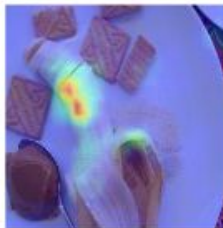
Predicted Answer : tennis

(a)



Question : What kind of bird is perched on the sill?

Predicted Answer : parrot



Question : What type of fruit is on the plate?

Predicted Answer : banana

(b)

Discussion

- Explanation Model not well motivated
 - Why counter examples a good explanation?
 - Maybe good for Yes/No questions.
 - How well it would apply to out of distribution images?
 - Distance based counterexamples alone are very close in performance to their proposed approach.
- How well the models are using images is also of concern?
 - Maybe some augmentation based on masking of train images would have been better for models to learn salient portions of an image.

How many doughnuts have sprinkles?

3

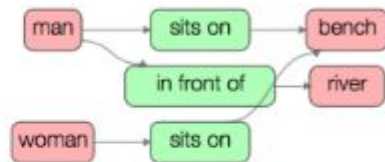
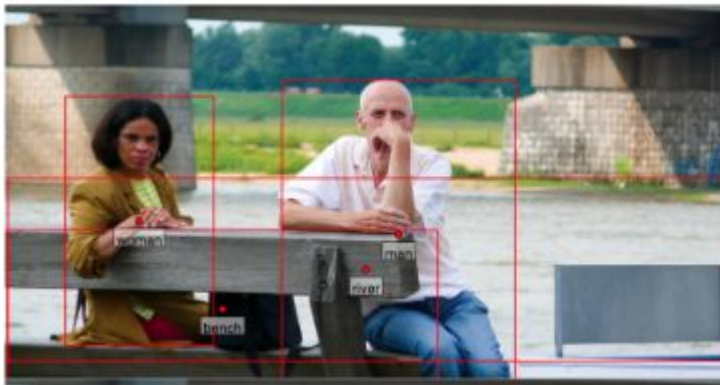


2



Discussion

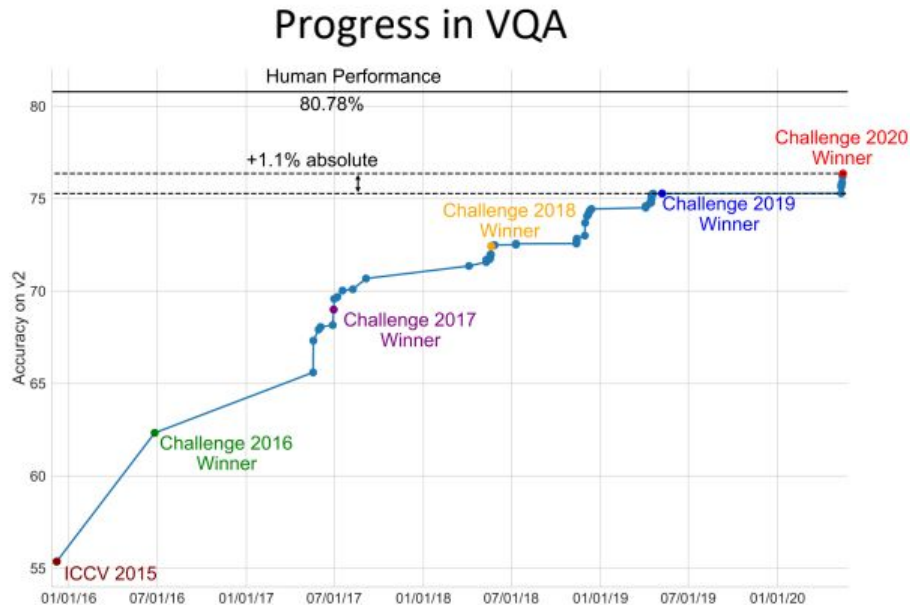
- VQA dataset covers surface level grounding task only.
 - The model does not learn how different objects interact in the image.
 - [Visual Genome](#), Krishna et. al. introduces a dataset which embeds visual constructs in language and relation between them using KGs.



A man and a woman sit on a park bench along a river.

Current Status

- [VQA Challenge](#), 2020.
- Approaches used:
 - Graph Neural Networks [[DL-61](#)]
 - Neural Architecture Search [[Yu et al.](#)]
 - Adversarial Training [[Gan et al.](#)]
 - Visual-Linguistic Transformers [[Bhargava et al.](#)]



Current Status

- [TextVQA](#)

- TextVQA requires models to **read and reason about text in an image** to answer questions based on them.



What does it say near the star on the tail of the plane?

Ground Truth

jet

Prediction

nothing

(a)



What is the time on bottom middle phone?

Ground Truth

15:20

Prediction

12:00

(b)

- [TextCaps](#)

- TextCaps requires models to **read and reason about text in images to generate captions** about them.



a

Model: a macdonald 's sign that is on a brick wall

Human: A tile wall with a red circle on it reading Mornington Crescent



b

Model: a sign that has the time of 12 : 37 on it

Human: A kiosk of track 13 of Metra which states that the 5:43 train has moved tracks

Current Status

VizWiz

- This task focuses on **answering visual questions** where **blind people were submitting images with recorded spoken questions**.



Q: Please can you tell me what this item is?
A: butternut squash red pepper soup



Q: Is it sunny outside?
A: yes



Q: Is this air conditioner on fan, dehumidifier, or air conditioning?
A: air conditioning

- Visual Dialog
 - Given an **image**, a **dialog history**, the agent has to **answer a follow-up question** in the dialog.



C: A dog with goggles is in a motorcycle side car.
Q: Is motorcycle moving or still?
A: It's parked
Q: What kind of dog is it?
A: Looks like beautiful pit bull mix
Q: What color is it?

Image

Dialog history

Question



Visual Dialog model

Answer

A: Light tan with white patch that runs up to bottom of his chin

Current Status

KnowIT VQA

- This task focuses on **answering questions** requiring understanding of **temporal, visual and textual modalities**.



Leonard: Have you noticed that Howard can take any topic and use it to remind you that he went to space?

Sheldon: Interesting hypothesis. Let's apply the scientific method.

Leonard: Okay. Hey, Howard, any thoughts on where we should get dinner?

Howard: Anywhere but the Space Station. On a good day, dinner was a bag full of meat loaf. But, hey, you don't go there for the food, you go there for the view.

Visual: How many people are there wearing glasses? *One*

Textual: Who has been to the space? *Howard*

Temporal: How do they finish the conversation? *Shaking hands*

Knowledge: Who owns the place where they are standing? *Stuart*



Questions?

Paper 2: Generating Natural Questions About an Image

Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, Lucy Vanderwende

ACL 2016

Image Captioning

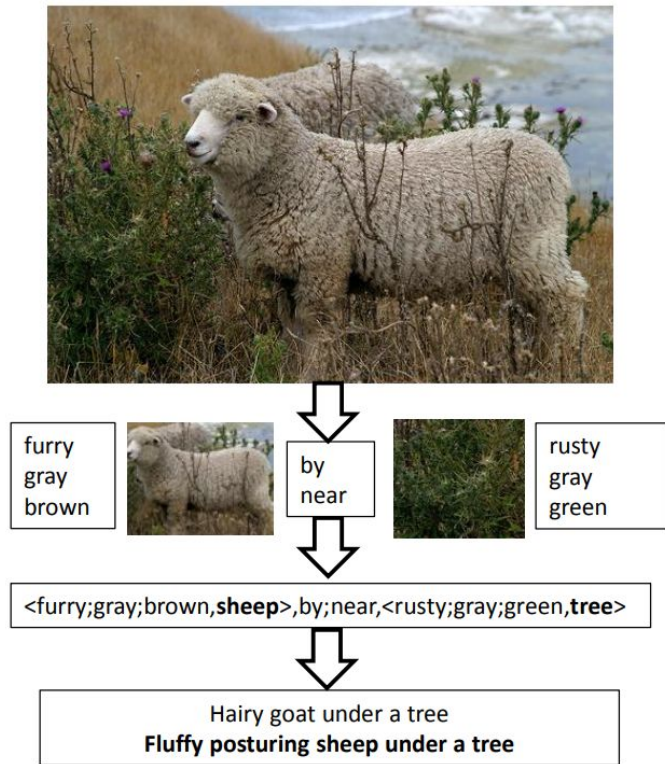
- Earliest works in image captioning tries to match words with images and subimages ([Barnard et al., 2002](#); [Barnard et al., 2001](#); [Mori et al., 2000](#)).



From Barnard et al., 2002

Improving Captioning

- [Li et al., 2011](#) extracted objects, visual attributes, and spatial attributes and put them into templated phrases, then combined these phrases for a description.
- [Yagcioglu et al., 2015](#) presents a retrieval method to gather descriptions from similar images.
- Recent methods popularize generating captions through recurrent and transformer language models ([Wang et al., 2020](#)).



From Li et al., 2011

Improving Captioning

- [Xu et al., 2015](#) used visual attention for better decoding.

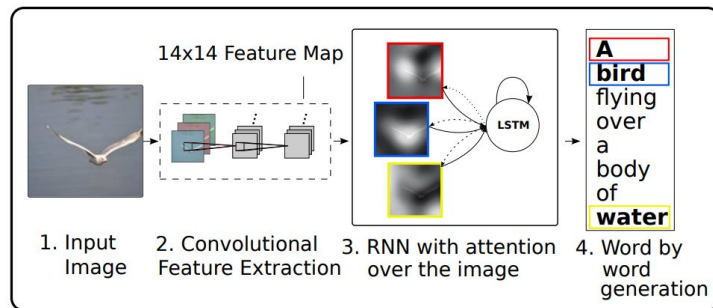
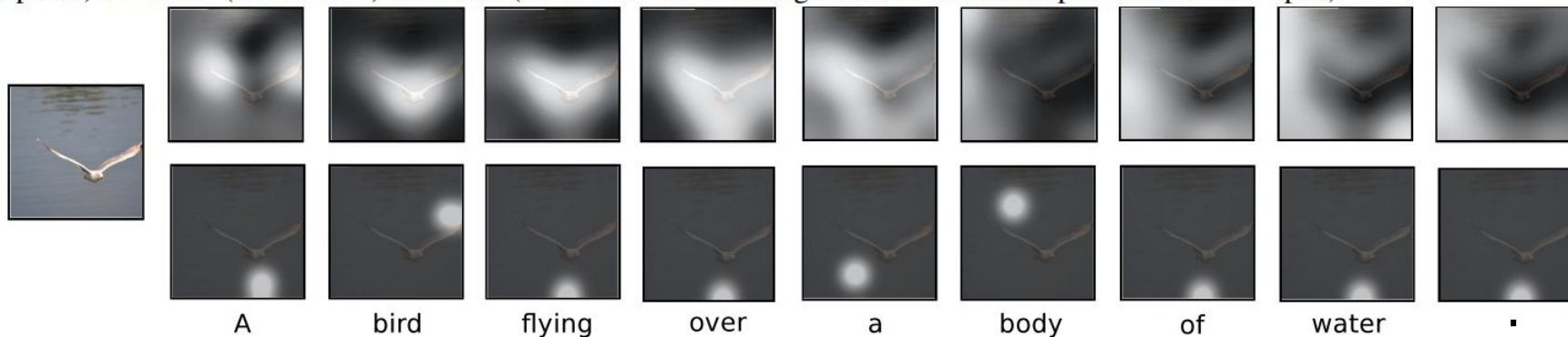


Figure 2. Attention over time. As the model generates each word, its attention changes to reflect the relevant parts of the image. “soft” (top row) vs “hard” (bottom row) attention. (Note that both models generated the same captions in this example.)



More Than a Literal Description

- [Chen et al., 2015](#) created tasks for creative captioning and paraphrasing.













Creative Image Captioning		Creative Visual Paraphrasing	
< Good >	< Bad >	< Good >	< Bad >
 <ul style="list-style-type: none"> -Hood under a full moon (*) -Mirror, mirror on the lake 	 <ul style="list-style-type: none"> -Falling water(*) -Can you see the dogs 	 <ul style="list-style-type: none"> -Bee on orange flowers(*) -When the flower looms, the bees come uninvited 	 <ul style="list-style-type: none"> -long haired girl(*) -Diamonds are a girl's best friend
 <ul style="list-style-type: none"> -Sky on the way home(*) -Red sky at night, Shepherd's delight 	 <ul style="list-style-type: none"> -City of lights (*) -Great balls of fire 	 <ul style="list-style-type: none"> -Lights in cave(*) -There is a light that never goes out 	 <ul style="list-style-type: none"> -Young roe deer(*) -The tree that looks like a deer
 <ul style="list-style-type: none"> -Sail on by (*) -Row, row, row your boat gently down the stream 	 <ul style="list-style-type: none"> -Red Bean Pastries (*) -When life gives you lemons 	 <ul style="list-style-type: none"> -Sky on the way home(*) -Go home, sky, you're drunk 	 <ul style="list-style-type: none"> -The flight of the crane(*) -That's a crane

Figure 5: Examples of creative captioning and creative visual paraphrasing. The left column shows good examples in blue, and the right column shows bad examples in red. The captions marked with * are the original captions of the corresponding query images.

More Than a Literal Description

- [Malinowski and Fritz, 2014](#) demonstrates a machine understanding of an image by question answering.



QA: (what is behind the table?, sofa)

Spatial relations exhibit different reference frames. Some annotations use observer-centric, others object-centric view

QA: (how many lights are on?, 6)

Moreover, some questions require detection of states 'light on or off'

More Than a Literal Description

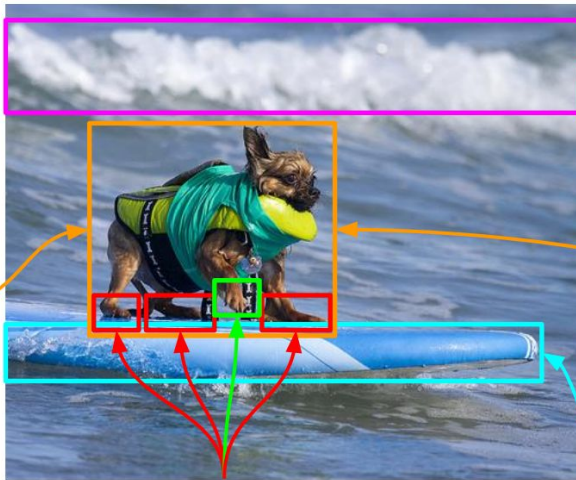
- [Zhu et al. 2016](#) proposes a another dataset for testing a machine's ability to look at and reason about specific regions in images by asking 7Ws.

Where does this scene take place?

- A) In the sea. ✓
- B) In the desert.
- C) In the forest.
- D) On a lawn.

What is the dog doing?

- A) Surfing. ✓
- B) Sleeping.
- C) Running.
- D) Eating.



Why is there foam?

- A) Because of a wave. ✓
- B) Because of a boat.
- C) Because of a fire.
- D) Because of a leak.

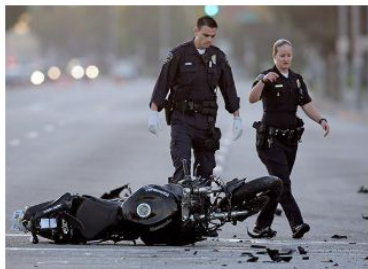
What is the dog standing on?

- A) On a surfboard. ✓
- B) On a table.
- C) On a garage.
- D) On a ball.

Which paw is lifted?

More Than a Literal Description

- Deciding on what to ask also demonstrate good understanding of an image.
- Visual Question Generation Task: **a system should generate a natural question about the image which can potentially engage a human in starting a conversation.**



Natural Questions:

- Was anyone injured in the crash?
- Is the motorcyclist alive?
- What caused this accident?

Generated Caption:

- A man standing next to a motorcycle.

Figure 1: Example image along with its natural questions and automatically generated caption.

Data Collection

Visual Question Generation

- Generate a natural question which can engage a human in starting a conversation.
- Crowdsourced on Amazon Mechanical Turk (AMT).
- Sourced from MS COCO, Flickr, and querying an image search engine (Bing).
- 5000 images from each data source, with 5 questions per image.
- Prompt is successful at capturing non-literal questions.



- How many horses are in the field? ❌
- Who won the race? ✅

Figure 2: Example right and wrong questions for the task of VQG.

Dataset Statistics

- 15,000 images with 75,000 questions.
- Average question length is 6 tokens.

# all images	15,000
# questions per image	5
# all workers participated	308
Max # questions written by one worker	6,368
Average work time per worker (sec)	106.5
Median work time per worker (sec)	23.0
Average payment per question (cents)	6.0

Table 2: Statistics of crowdsourcing task, aggregating all three datasets.

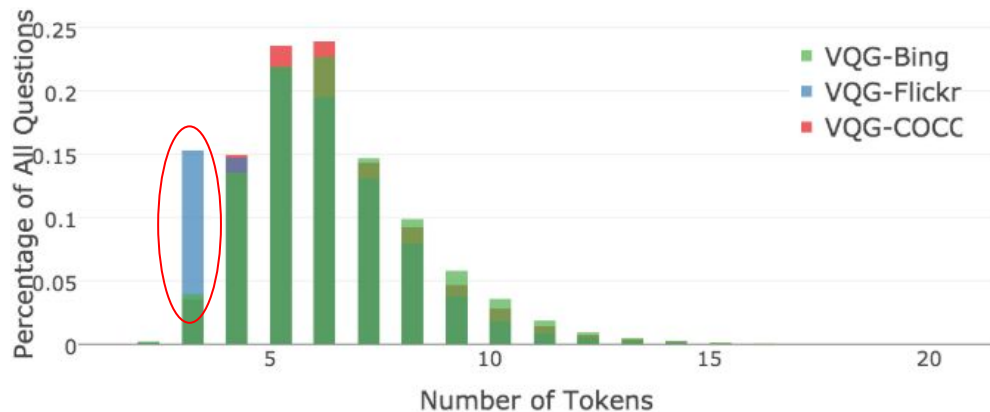


Figure 5: Average annotation length of the three VQG datasets.

- Most questions begin with 'what', 'is', and 'how'.

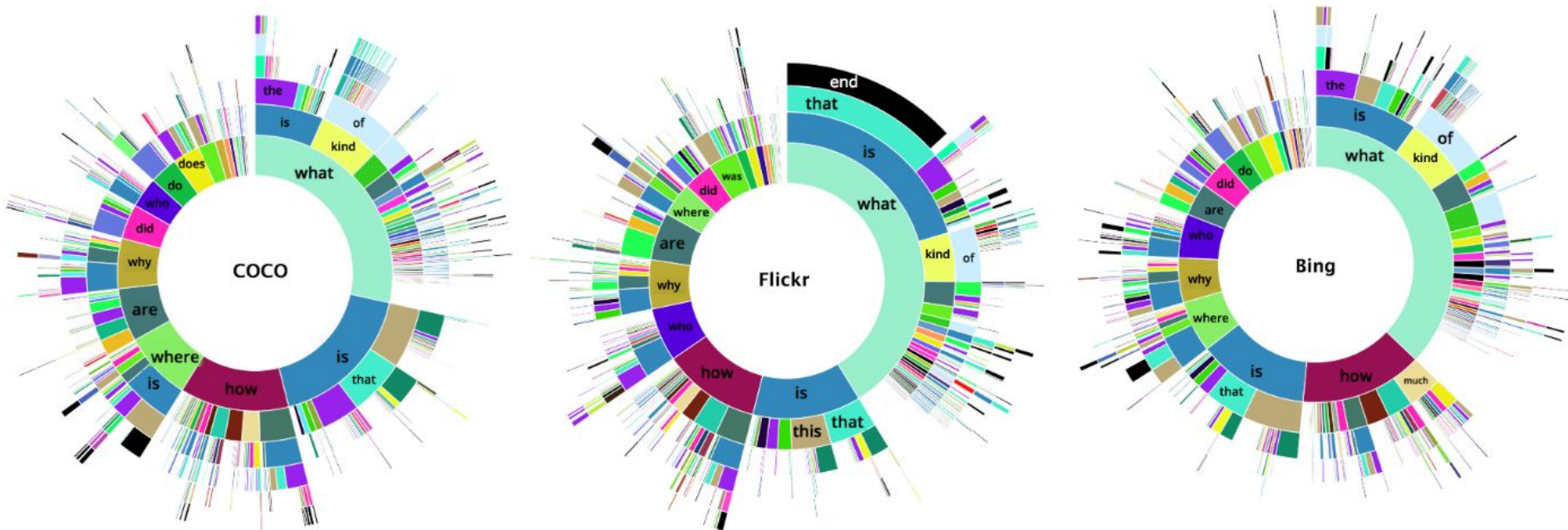


Figure 6: VQG N-gram sequences. 'End' token distinguishes natural ending with n-gram cut-off.

VQG COCO-5000

- Sampled 5000 images of MS COCO which are also annotated by the CQA dataset ([Ren et al., 2015](#)) and by VQA ([Antol et al., 2015](#)).
- Object-centric questions.
- CQA: Generation by rule application from captions. Not always coherent.
- VQA: Ask a question about this scene that a smart robot probably cannot answer, but any human can easily answer while looking at the scene in the image.



Dataset	Annotations
COCO	- A man holding a box with a large chocolate covered donut.
CQA	- What is the man holding with a large chocolate-covered doughnut in it?
VQA	- Is this a large doughnut? - Why is the donut so large? - Is that for a specific celebration?
VQG	- Have you ever eaten a donut that large before? - Is that a big donut or a cake? - Where did you get that?

Table 1: Dataset annotations on the above image.

- VQG mentions more of the objects in the images than CQA and VQA.
- VQG has a larger vocabulary than CQA and VQA, indicating greater diversity in question formulation.
- VQG usage of verb POS is comparable to VQA.
- VQG contains more abstract concepts.
- VQG has highest inter-annotator textual similarity, so there are some consensus in asking a natural question.

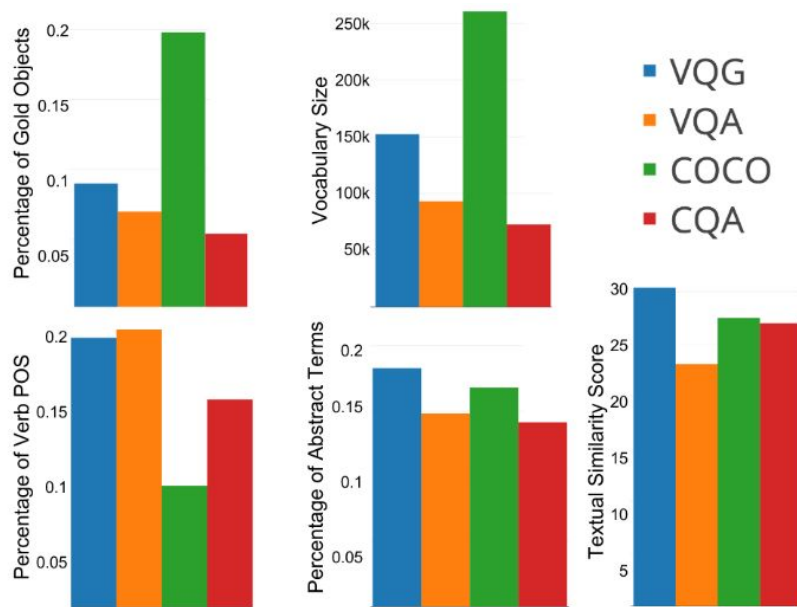


Figure 3: Comparison of various annotations on the MS COCO dataset. (a) Percentage of gold objects used in annotations. (b) Vocabulary size (c) Percentage of verb POS (d) Percentage of abstract terms (e) Inter-annotation textual similarity score.

VQG Flickr-5000

- Object words such as ‘cat’ and ‘dog’ are very frequent in the MS COCO dataset.
- This motivates the collection of another dataset.

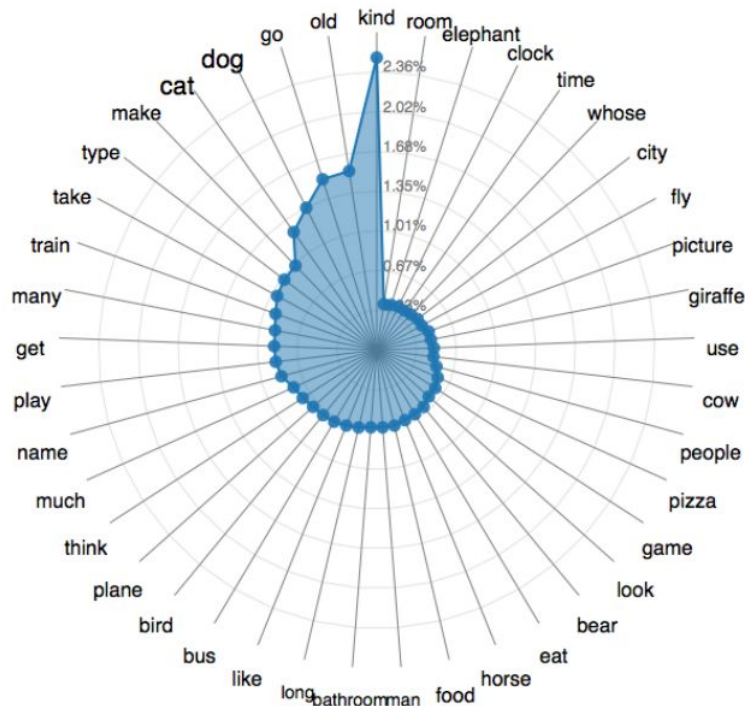
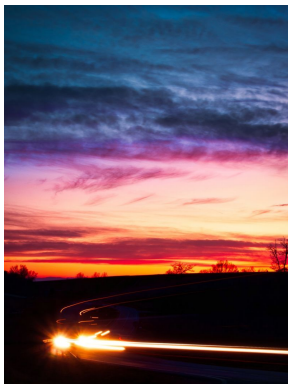


Figure 4: Frequency graph of top 40 words in $VQG_{COCO-5000}$ dataset.

VQG Bing-5000

- Select 1,200 most frequent event-centric words based on Project Gutenberg frequencies.
- Query Bing with each term. Take first 4 to 5 images retrieved that are not graphics nor cartoons.
- Event-centric questions.
- Substantially different from MS COCO dataset.



Captions Bing-5000

- Crowdsourced 5 captions for each image in the VQG Bing-5000 dataset using the same prompt as used to source the MS COCO captions.
- 25,000 gold captions.
- Gap in state-of-the-art performance indicates VQG Bing-5000 presents a new class of images.

BLEU		METEOR	
<i>Bing</i>	<i>MS COCO</i>	<i>Bing</i>	<i>MS COCO</i>
0.101	0.291	0.151	0.247

Table 3: Image captioning results

Models

Generative Models

- Use VGGNet ([Simonyan and Zisserman, 2014](#)) architecture to compute a 4096-dimensional output (from *fc7* layer) as deep convolutional image features.
- MELM ([Fang et al., 2015](#)) and MT model generate less coherent sentences.

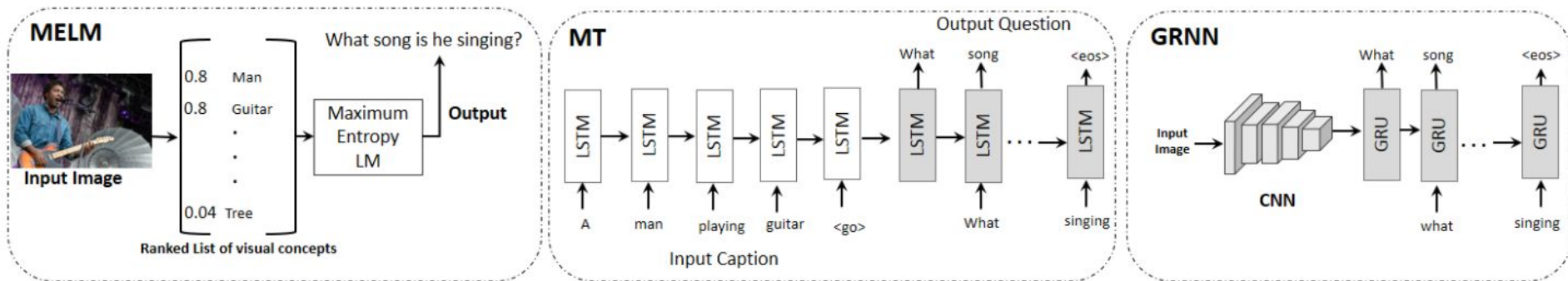


Figure 7: Three different generative models for tackling the task of VQG.

Retrieval Methods

- In Image Captioning, 80% of generated captions at test time (using Vinyals et al., 2015) were exactly identical to training set options.
- Use VGGNet fc7 output vector to compute the distance between images.
- One-best question: the question with the highest semantic similarity to other 4 questions (using BLEU).

- 1-NN: $K=1$, retrieves the closest image and emits its one-best question.
- K-NN + min: $K=30$, $\text{max-distance}=0.35$, and $\text{min-distance}=0.1$.
 - max-distance: a parameter as an upper bound distance for including an image in candidate pool. Avoid images that make the pool noisy.
 - min-distance: a parameter to set one very similar image as the only candidate image.
 - Emit question with highest Smoothed-BLEU and Average-Word2Vec (gensim) similarity with rest of pool.

Evaluation

Human Evaluation

- Asked 3 crowd workers on AMT to each rate the quality of candidate questions on a three-point semantic scale.
- 3 is considered best.

Automatic Evaluation

- BLEU metric up to 4-grams.
- METEOR: default setting on version 1.5.
- Δ BLEU ([Galley et al., 2015](#)): for evaluating tasks with diverse references.
 - Majority rating on scale of 1-3 from crowd-sourced 3 human ratings per reference.
- Pearson's r , Spearman's ρ , and Kendall's τ .

	METEOR	BLEU	ΔBLEU
r	0.916 (4.8e-27)	0.915 (4.6e-27)	0.915 (5.8e-27)
ρ	0.628 (1.5e-08)	0.67 (7.0e-10)	0.702 (5.0e-11)
τ	0.476 (1.6e-08)	0.51 (7.9e-10)	0.557 (3.5e-11)

Table 6: Correlations of automatic metrics against human judgments, with p-values in parentheses.

Results

- Randomly divided each VQG-5000 dataset into train (50%), val (25%), and test(25%).
- Each model is trained on all train dataset and also independently on each VQG-5000 train dataset.
- Human-Consensus annotation: same as one-best.
- Human-Random annotation: randomly chosen among 5 human annotations.

		<i>Human_{consensus}</i>	<i>Human_{random}</i>	<i>GRNN_X</i>	<i>GRNN_{all}</i>	<i>1-NN_{bleu-X}</i>	<i>1-NN_{gensim-X}</i>	<i>K-NN+min_{bleu-X}</i>	<i>K-NN+min_{gensim-X}</i>	<i>1-NN_{bleu-all}</i>	<i>1-NN_{gensim-all}</i>	<i>K-NN+min_{bleu-all}</i>	<i>K-NN+min_{gensim-all}</i>
Human Evaluation													
	Bing	2.49	2.38	1.35	1.76	1.72	1.72	1.69	1.57	1.72	1.73	1.75	1.58
	COCO	2.49	2.38	1.66	1.94	1.81	1.82	1.88	1.64	1.82	1.82	1.96	1.74
	Flickr	2.34	2.26	1.24	1.57	1.44	1.44	1.54	1.28	1.46	1.46	1.52	1.30
Automatic Evaluation													
BLEU	Bing	87.1	83.7	12.3	11.1	9.0	9.0	11.2	7.9	9.0	9.0	11.8	7.9
	COCO	86.0	83.5	13.9	14.2	11.0	11.0	19.1	11.5	10.7	10.7	19.2	11.2
	Flickr	84.4	83.6	9.9	9.9	7.4	7.4	10.9	5.9	7.6	7.6	11.7	5.8
MET.	Bing	62.2	58.8	16.2	15.8	14.7	14.7	15.4	14.7	14.7	14.7	15.5	14.7
	COCO	60.8	58.3	18.5	18.5	16.2	16.2	19.7	17.4	15.9	15.9	19.5	17.5
	Flickr	59.9	58.6	14.3	14.9	12.3	12.3	13.6	12.6	12.6	12.6	14.6	13.0
Δ BLEU	Bing	63.38	57.25	11.6	10.8	8.28	8.28	10.24	7.11	8.43	8.43	11.01	7.59
	COCO	60.81	56.79	12.45	12.46	9.85	9.85	16.14	9.96	9.78	9.78	16.29	9.96
	Flickr	62.37	57.34	9.36	9.55	6.47	6.47	9.49	5.37	6.73	6.73	9.8	5.26

		<i>Human_{consensus}</i>	<i>Human_{random}</i>	<i>GRNN_X</i>	<i>GRNN_{all}</i>	<i>I-NN_{bleu-X}</i>	<i>I-NN_{gensim-X}</i>	<i>K-NN+min_{bleu-X}</i>	<i>K-NN+min_{gensim-X}</i>	<i>I-NN_{bleu-all}</i>	<i>I-NN_{gensim-all}</i>	<i>K-NN+min_{bleu-all}</i>	<i>K-NN+min_{gensim-all}</i>
Human Evaluation													
	Bing	2.49	2.38	1.35	1.76	1.72	1.72	1.69	1.57	1.72	1.73	1.75	1.58
	COCO	2.49	2.38	1.66	1.94	1.81	1.82	1.88	1.64	1.82	1.82	1.96	1.74
	Flickr	2.34	2.26	1.24	1.57	1.44	1.44	1.54	1.28	1.46	1.46	1.52	1.30
Automatic Evaluation													
BLEU	Bing	87.1	83.7	12.3	11.1	9.0	9.0	11.2	7.9	9.0	9.0	11.8	7.9
	COCO	86.0	83.5	13.9	14.2	11.0	11.0	19.1	11.5	10.7	10.7	19.2	11.2
	Flickr	84.4	83.6	9.9	9.9	7.4	7.4	10.9	5.9	7.6	7.6	11.7	5.8
MET.	Bing	62.2	58.8	16.2	15.8	14.7	14.7	15.4	14.7	14.7	14.7	15.5	14.7
	COCO	60.8	58.3	18.5	18.5	16.2	16.2	19.7	17.4	15.9	15.9	19.5	17.5
	Flickr	59.9	58.6	14.3	14.9	12.3	12.3	13.6	12.6	12.6	12.6	14.6	13.0
Δ BLEU	Bing	63.38	57.25	11.6	10.8	8.28	8.28	10.24	7.11	8.43	8.43	11.01	7.59
	COCO	60.81	56.79	12.45	12.46	9.85	9.85	16.14	9.96	9.78	9.78	16.29	9.96
	Flickr	62.37	57.34	9.36	9.55	6.47	6.47	9.49	5.37	6.73	6.73	9.8	5.26

		<i>Human_{consensus}</i>	<i>Human_{random}</i>	<i>GRNN_X</i>	<i>GRNN_{all}</i>	<i>1-NN_{bleu-X}</i>	<i>1-NN_{gensim-X}</i>	<i>K-NN+min_{bleu-X}</i>	<i>K-NN+min_{gensim-X}</i>	<i>1-NN_{bleu-all}</i>	<i>1-NN_{gensim-all}</i>	<i>K-NN+min_{bleu-all}</i>	<i>K-NN+min_{gensim-all}</i>
Human Evaluation													
	Bing	2.49	2.38	1.35	1.76	1.72	1.72	1.69	1.57	1.72	1.73	1.75	1.58
	COCO	2.49	2.38	1.66	1.94	1.81	1.82	1.88	1.64	1.82	1.82	1.96	1.74
	Flickr	2.34	2.26	1.24	1.57	1.44	1.44	1.54	1.28	1.46	1.46	1.52	1.30
Automatic Evaluation													
BLEU	Bing	87.1	83.7	12.3	11.1	9.0	9.0	11.2	7.9	9.0	9.0	11.8	7.9
	COCO	86.0	83.5	13.9	14.2	11.0	11.0	19.1	11.5	10.7	10.7	19.2	11.2
	Flickr	84.4	83.6	9.9	9.9	7.4	7.4	10.9	5.9	7.6	7.6	11.7	5.8
MET.	Bing	62.2	58.8	16.2	15.8	14.7	14.7	15.4	14.7	14.7	14.7	15.5	14.7
	COCO	60.8	58.3	18.5	18.5	16.2	16.2	19.7	17.4	15.9	15.9	19.5	17.5
	Flickr	59.9	58.6	14.3	14.9	12.3	12.3	13.6	12.6	12.6	12.6	14.6	13.0
Δ BLEU	Bing	63.38	57.25	11.6	10.8	8.28	8.28	10.24	7.11	8.43	8.43	11.01	7.59
	COCO	60.81	56.79	12.45	12.46	9.85	9.85	16.14	9.96	9.78	9.78	16.29	9.96
	Flickr	62.37	57.34	9.36	9.55	6.47	6.47	9.49	5.37	6.73	6.73	9.8	5.26

Discussion


Q. Explosion

Hurricane

Rain Cloud

Car Accident

Human

- What caused this explosion?
- Was this explosion an accident?

- What caused the damage to this city?
- What happened to this place?

- Are those rain clouds?
- Did it rain?

- Did the drivers of this accident live through it?
- How fast were they going?

GRNN

- How much did the fire cost?
- What is being burned here?

- What happened to the city?
- What caused the fall?

- What kind of clouds are these?
- Was there a bad storm?

- How did the car crash?
- What happened to the trailer?

KNN

- What caused this fire?

- What state was this earthquake in?

- Did it rain?

- Was anybody hurt in this accident?

Caption

- A train with smoke coming from it.

- A pile of dirt.

- Some clouds in a cloudy day.

- A man standing next to a motorcycle.



Human

- How long did it take to make that ice sculpture?

- Is the dog looking to take a shower?

GRNN

- How long has he been hiking?

- Is this in a hotel room?

KNN

- How deep was the snow?

- Do you enjoy the light in this bathroom?

Discussion

- Created 3 datasets which covers a range of object-centric to event-centric images.
- **A system should generate a natural question about the image which can potentially engage a human in starting a conversation.**
 - Do you think the definition of a question in this task is vague?
 - Do you want to have a discussion with the sample questions?

Discussion

- Retrieval system uses similarity between images to select a question. On VQG COCO-5000 and Flickr-5000, the retrieval system outperforms the neural system.
 - Why is there a need for a retrieval system?
 - Can the retrieval system be a good baseline for the amount of context a neural system should learn to generate questions?

Discussion

- Curated gold captions for the event-centric dataset and showed that the dataset challenges the state-of-the-art image captioning models.
- Performed analysis to conclude that end-to-end deep neural models outperform other approaches on the challenging event-centric dataset.
 - How the difficulty of retrieval models to generalize to VGQ Bing-5000 indicate that VGQ Bing-5000 is a challenging dataset?

Discussion

- Showed that the automatic metric Δ BLEU strongly correlates with human judgements for task evaluation.
- BLEU can be a good proxy if Δ BLEU cannot be made.
 - The authors did not present a criteria for the human evaluation rating scale (1-3). Can we trust this claim?
 - Should we use BLEU to evaluate how much context the machine knows about the image through the question?

Discussion

- We learned previously from Neural Text DeGeneration ([Holtzman et al., 2019](#)) that natural language does not maximize probability.
 - Does checking the generated question matches one of the 5 reference questions strongly in terms of BLEU indicate good question generation?

Discussion

- The author proposes to use VQG to improve context learning in dialogue systems.
- Despite the intrinsic relationship between questioning and answering, VQG and VQA are usually explored separately in literature. Recent research proposes to combine the two ([Tang et al. 2018](#); [Li et al. 2017](#)) to improve models on both task at the same time.
- VQG has also been applied to other areas of work for models to gain more understanding from images in new fields. Radiology ([Sarrouti et al. 2020](#)) and Art ([Garcia et al., 2020](#)) are examples of this.
- What are some other possible extensions of this work?



Questions?