ALLEN WU
COS 598B, Spring ‘18

VQA Method: Simple Baselines

Roadmap

1. Overview

2. Simple Baseline for VQA

3. What's in a question?: Using visual questions as a Form of Supervision

1.

OVERVIEW

Visual Question Answering

"We have built a smart robot. It understands a lot about images. It can recognize and name all the objects, it knows where the objects are, it can recognize the scene (e.g., kitchen, beach), people's expressions and poses, and properties of objects (e.g., color of objects, their texture).
Your task is to stump this smart robot!
Ask a question about this scene that this smart robot probably can not answer, but any human can easily answer while looking at the scene in the image."

Visual Question Answering

- Combining NLP with CV for high-level scene interpretation
- Emergence of large image datasets and text corpus
- Image Captioning ⟶ Visual Question Answering (harder)
  - Wider range of knowledge, more reasoning skills
  - Easiest Description ⟶ Correct Answer
    In VQA, the answer is clearer, artificial metrics not necessary

Visual Question Answering

- New mode of interaction
- It is to become a natural human-AI interaction paradigm
- “What breed of dog is this?”
  - Pretty natural setting
  - Information revelation
- “Why is he doing that?”
  - unusual or unexpected stuff going on…

2.

Simple Baseline for Visual Question Answering

- Fully explored the potential of simple Bag-of-Words baseline achieves RNN-comparable performance
- Good Web Demo and separation of word and image score ranking help us understand what has been learned by the machine
- Informative image regions highlight using CAM

Simple Baseline for Visual Question Answering

By: Bolei Zhou et al., 2015

VQA and the model-to-be-modified

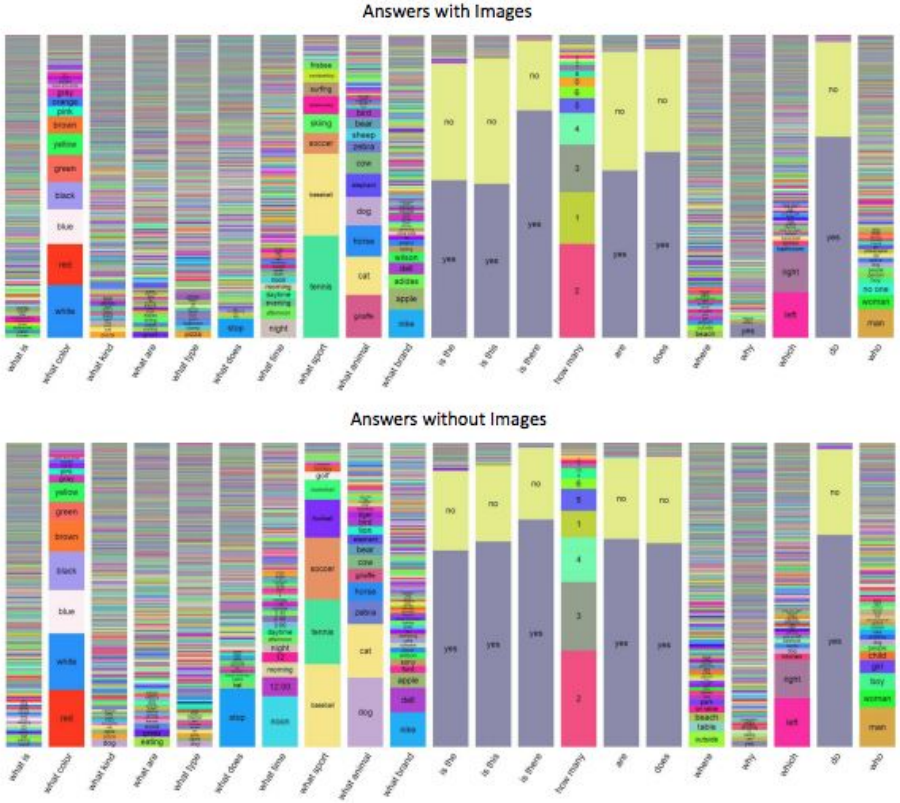
- All kinds of questions
  - “What books are under the television?”(image itself)
  - “Which chair is the most expensive?”(knowledge beyond the image content)
- Previous BOWIMG
  - Bag-of-words+image feature outperforms the LSTM-based models on a synthesized visual QA dataset on top of image captions of COCO
  - Larger COCO VQA dataset, not so good

More of old model: Bow Q + CNN features

- Bag of words Question(Bow Q): The top 1,000 words in the questions & top 10 first, second and third words of the questions (1030-dim embedding)

- CNN features: The activations from the last hidden layer of VGGNet (4096-dim embedding)

Correlation of starting words and answers from Fig 5 of VQA paper



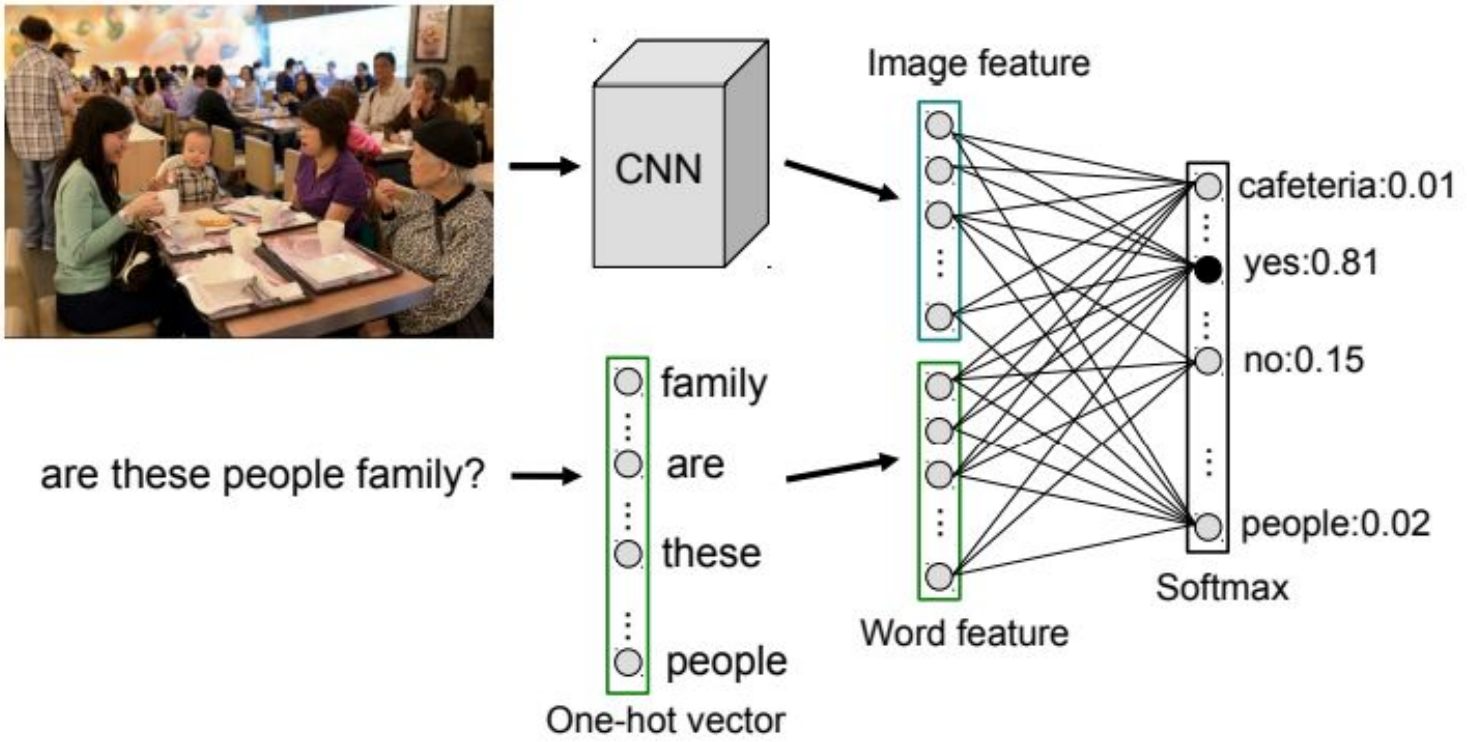
Old BoW vs LSTM Q

	Open-Ended				Multiple-Choice			
	All	Yes/No	Number	Other	All	Yes/No	Number	Other
BoW Q + I	52.64	75.55	33.67	37.37	58.97	75.59	34.35	50.33
deeper LSTM Q + norm I	**57.75**	**80.50**	**36.77**	**43.08**	**62.70**	**80.52**	**38.22**	**53.01**

Simple Baseline for VQA: Overview

- Most models:
  - Classification task
  - Visual feature: VGG or GoogLeNet
  - Word features: LSTM-based features
- iBOWIMG:
  - Simple bag-of-words as text feature
  - BOW features from Questions + GoogLeNet deep visual features from the image
  - Achieves comparable performance to RNN approaches

iBOWIMG Structure



Word feature+Visual feature

$$r = \mathbf{M}_w \mathbf{x}_w + \mathbf{M}_v \mathbf{x}_v .$$

Question: what is the color of the sofa
Predictions:

brown (score: 12.89 = 1.01 [image] + 11.88 [word])
red (score: 11.92 = 1.13 [image] + 10.79 [word])
yellow (score: 11.91 = 1.08 [image] + 10.84 [word])

Based on image only: books (3.15), yes (3.14), no (2.95)
Based on word only: brown (11.88), gray (11.18), tan (11.16)

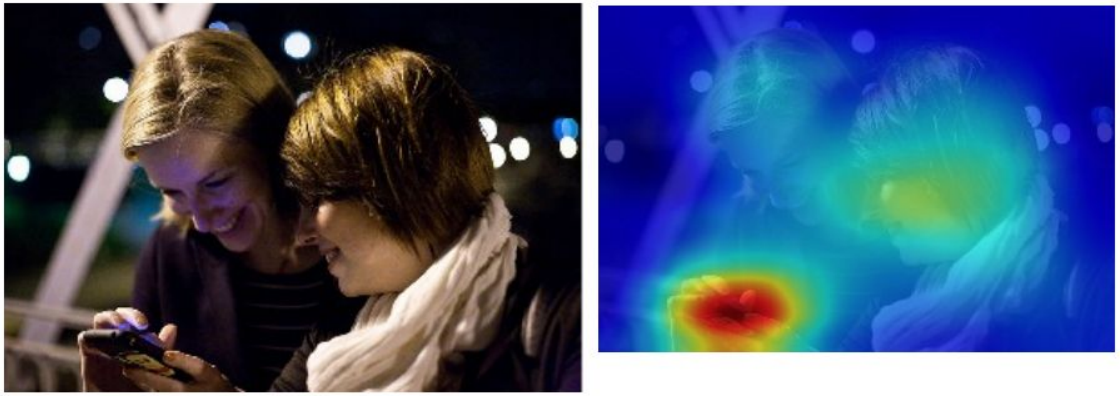
Weight ranking and comparison

Question: what is the color of the sofa
Predictions:
brown (score: 12.89 = 1.01 [image] + 11.88 [word])
red (score: 11.92 = 1.13 [image] + 10.79 [word])
yellow (score: 11.91 = 1.08 [image] + 10.84 [word])

Based on image only: books (3.15), yes (3.14), no (2.95)
Based on word only: brown (11.88), gray (11.18), tan (11.16)

- Here, most of the weights come from question words, the bias in the frequency of object and actions appearing in the images of COCO dataset

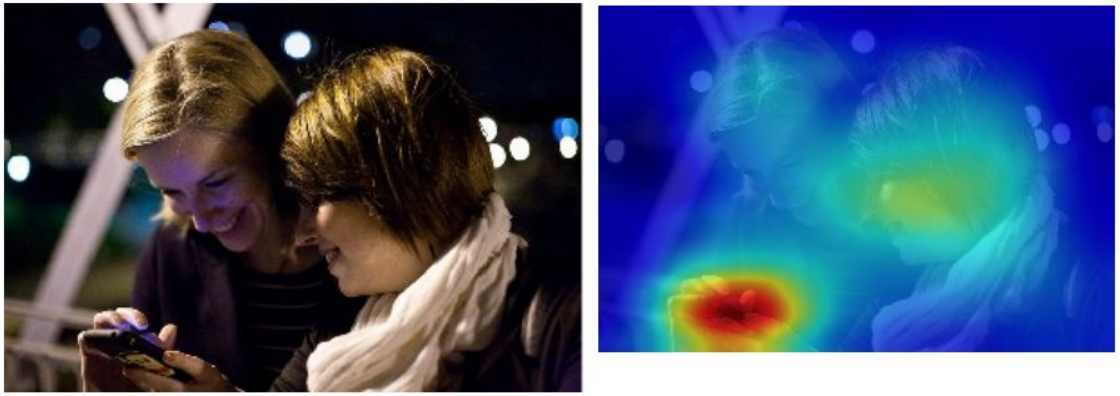
Word importance in Questions



Question: What are they doing?
Prediction: texting (score: 12.02=3.78 [image] + 8.24 [word])
Word importance: doing(7.01) are(1.05) they(0.49) what(-0.3)

- So we can just add up the column of weight matrix to get the importance of each word in the question
- 8.24=7.01+1.05+0.49-0.3

Implicit image attention from last Conv layer



Question: What are they doing?
Prediction: texting (score: 12.02=3.78 [image] + 8.24 [word])
Word importance: doing(7.01) are(1.05) they(0.49) what(-0.3)

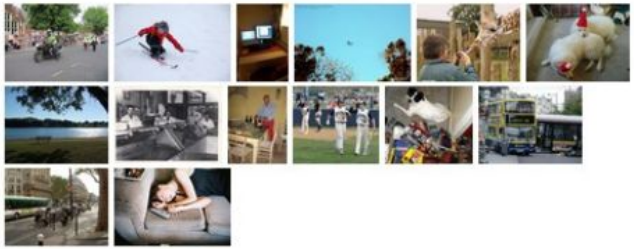
- Informative image regions relevant to the predicted answers using Class Activation mapping (CAM)
- In the picture, cellphone is highlighted as people are texting

Training details

- Learning rate and weight clip
  - Different learning rate and weight clipping for the word embedding layer(much higher) and softmax layer
- Tuning model parameters(manually of course)
  - # epochs to train
  - Learning rate and weight clip
  - Threshold for removing less frequent question word and answer classes

Online Demo

1.Click One:



2.Type Question: what are people doing

Question: what are people doing

Predictions::

- **flying kites** (score: 12.86 = 1.64 [image] + 11.22 [word])
- **playing baseball** (score: 12.38 = 3.18 [image] + 9.20 [word])
- **playing frisbee** (score: 11.96 = 1.72 [image] + 10.24 [word])

Based on image only: baseball (4.74), batting (4.44), glove (4.12),
Based on word only: playing wii (11.49), flying kites (11.22), playing frisbee (10.24),

Question: where is the place

Predictions::

- **field** (score: 10.63 = 3.05 [image] + 7.58 [word])
- **park** (score: 9.69 = 2.96 [image] + 6.73 [word])
- **in air** (score: 9.67 = 2.27 [image] + 7.40 [word])

Based on image only: baseball (4.74), batting (4.44), glove (4.12),
Based on word only: above stove (8.23), behind clouds (8.08), on floor (8.03),

Online Demo: my own test

Question: What is she wearing

Predictions::

- **yes** (score: 11.86 = 4.54 [image] + 7.32 [word])
- **no** (score: 11.25 = 4.19 [image] + 7.07 [word])
- **bikini** (score: 7.75 = 1.97 [image] + 5.78 [word])

Based on image only: kite (5.04), sunset (5.02), yes (4.54),
Based on word only: yes (7.32), no (7.07), capris (6.22),

Experiments

- COCO VQA dataset
- **Standard splits**: 248,349 pairs in train2014 and 121,512 pairs in val2014, for 123,287 images overall in the training set
- 3 questions annotated for each image in COCO
- **Majority voting** on 10 ground-truth answers for each question
- 3 Q-A pairs from each image for training
- val2014=70%A+30%B
- train2014+A are for training, B is used as validation set for parameter tuning

Experiments

- After finding the best model parameters, train2014+val2014 are used to train, tested on COCO test2015
- Open-Ended Question : top-1 predicted answer from the softmax output
- Multiple-Choice Question : get the softmax probability for each of the given choices then select the most confident one

Test-dev dataset (unlimited submission)

Table 1: Performance comparison on test-dev.

	Open-Ended				Multiple-Choice			
	Overall	yes/no	number	others	**Overall**	yes/no	number	others
IMG [2]	28.13	64.01	00.42	03.77	30.53	69.87	00.45	03.76
BOW [2]	48.09	75.66	36.70	27.14	53.68	75.71	37.05	38.64
BOWIMG [2]	52.64	75.55	33.67	37.37	58.97	75.59	34.35	50.33
LSTMIMG [2]	53.74	78.94	35.24	36.42	57.17	78.95	35.80	43.41
CompMem [6]	52.62	78.33	35.93	34.46	-	-	-	-
NMN+LSTM [1]	54.80	77.70	37.20	39.30	-	-	-	-
WR Sel. [13]	-	-	-	-	60.96	-	-	-
ACK [16]	55.72	79.23	36.13	40.08	-	-	-	-
DPPnet [11]	**57.22**	80.71	37.24	41.69	**62.48**	80.79	38.94	52.16
iBOWIMG	55.72	76.55	35.03	42.62	61.68	76.68	37.05	54.44

Test-dev dataset (unlimited submission)

Table 1: Performance comparison on test-dev.

	Open-Ended				Multiple-Choice			
	Overall	yes/no	number	others	**Overall**	yes/no	number	others
IMG [2]	28.13	64.01	00.42	03.77	30.53	69.87	00.45	03.76
BOW [2]	48.09	75.66	36.70	27.14	53.68	75.71	37.05	38.64
BOWIMG [2]	52.64	75.55	33.67	37.37	58.97	75.59	34.35	50.33
LSTMIMG [2]	53.74	78.94	35.24	36.42	57.17	78.95	35.80	43.41
CompMem [6]	52.62	78.33	35.93	34.46	-	-	-	-
NMN+LSTM [1]	54.80	77.70	37.20	39.30	-	-	-	-
WR Sel. [13]	-	-	-	-	60.96	-	-	-
ACK [16]	55.72	79.23	36.13	40.08	-	-	-	-
DPPnet [11]	**57.22**	80.71	37.24	41.69	**62.48**	80.79	38.94	52.16
iBOWIMG	55.72	76.55	35.03	42.62	61.68	76.68	37.05	54.44

Test-standard dataset (limited submission)

Table 2: Performance comparison on test-standard.

	Open-Ended				Multiple-Choice			
	Overall	yes/no	number	others	**Overall**	yes/no	number	others
LSTMIMG [2]	54.06	-	-	-	-	-	-	-
NMN+LSTM [1]	55.10	-	-	-	-	-	-	-
ACK [16]	55.98	79.05	36.10	40.61	-	-	-	-
DPPnet [11]	**57.36**	80.28	36.92	42.24	**62.69**	80.35	38.79	52.79
iBOWIMG	55.89	76.76	34.98	42.62	61.97	76.86	37.30	54.60

Simple Baseline for VQA: **Conclusion**

- A simple baseline achieves comparable performance to several RNN-based approaches
- Move beyond from memorizing the correlations to actual reasoning and understanding of the question and image
- Why Bag-Of-Words is sufficient?
  - RNN has learned a little
  - Bad sign of effective learning
  - They are all kind of bad ~ 60%

3.

What's in a Question?

- Leveraging correlation between questions
- Generated more training examples
- Able to utilize information of unanswered questions even at test time

Question: Have any researchers taken more advantage of these kinds of question information retrieval?

What's in a Question: Using Visual Questions as a Form of Supervision

By: Siddha Ganju, Olga Russakovsky, Abhinav Gupta CVPR 2017

Questions have information

- Collecting fully annotated image datasets is challenging and expensive
- Key observation : question itself provides useful information about the image
  - “Are **people** waiting for the food **truck**?” indicates the presence of people and truck
  - “How many **umbrellas** are in the image?” indicates the presence of umbrella
- Not single-sided interaction with humans soliciting information from AI systems anymore, but a form of supervision to improve computer vision systems

Reasoning from a Question

- that animal must be a dog
- breed must be a property of dog
- all dogs in this picture must be the same breed
- knowing the breed must be important

Correlation between questions and image

- COCO dataset
  - 3 Visual questions
  - 5 Image captions
  - Image classification labels (80 target object classes)
- Two perspectives of examining the information content of the questions
  - Can questions provide a good image description?
  - Can we learn what objects are present in the image, given questions?

Can questions provide a good image description

- People sometimes are compelled to ask a question “Are the flowers real or artificial?”
- Quantitative results:
  - METEOR and SPICE
  - One Q: use one of the visual questions directly as caption
  - Three Qs: use concatenated all three questions as caption

Can questions provide a good image description

- Can questions provide a good image description?
  - People sometimes are compelled to ask a question “Are the flowers real or artificial?”
  - Quantitative results:
    - Seq2Seq: a model trained on COCO taking an input of three visual Qs and output a semantically meaningful image caption
    - NT: a computer vision model that takes in an image and outputs an image caption
    - NT+Seq2Seq: Vanilla concatenation

Seq2Seq: 3Qs->image caption

What are these two people doing in the scene? What color is the person on the right's hat? Was this picture taken during the day?
people during day with hat

Is this rice noodle soup? What is to the right of the soup? What website copyrighted the picture?
copyrighted noodle soup

Correlation: Between Qs and Q+I

Information	Model	METEOR	SPICE
Qs-only	One Q	0.089	0.058
	Three Qs	0.140	0.115
	Seq2Seq	0.206	0.140
Image-only	NT [25]	0.267	0.194
Image+Qs	NT + Seq2Seq	0.305	0.256

- (3Qs outperforms 1Q) different questions provide complementary information about the image content
- (NT+Seq2Seq outperforms NT) Signal from visual questions may be complementary to the information in the image

What objects are present by Qs?

64 question types in COCO

What is the	What is	Is the	Is this	Is this a	Is there a	Is it	Is there	Is	Is this an	Is that a

Table 6: Unconfirmed Question Types

How many	What	What color is the	Are the	What kind of
What type of	What are the	Where is the	Does the	What color are the
Are these	Are there	Which	What is the man	Are
How	Does this	What is on the	What does the	How many people are
What is in the	What is this	Do	What are	Are they
What time	What sport is	Are there any	What color is	Why
Where are the	What color	Who is	What animal is	Do you
How many people are in	What room is	Has	What is the woman	Can you
Why is the	What is the color of the	What is the person	Could	Was
What number is	What is the name	What brand	Is the person	Is he
Is the man	Is the woman	Is this person		

Table 7: Confirmed Question Types

Object classification

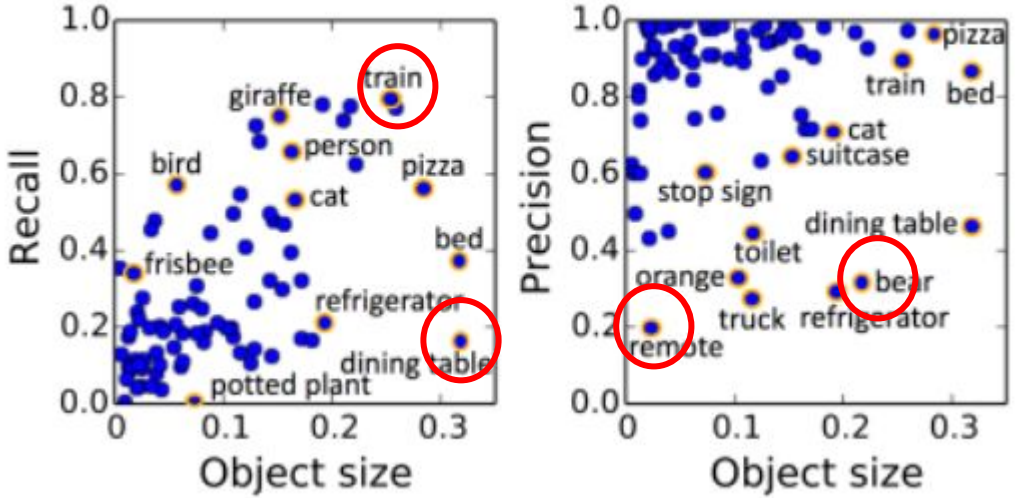
- Different Q type
  - “How many different **flowers** are on the **table**?”
  - “Is there a zebra in the photo?”
- Extract objects on 80 COCO classes
  - NLTK to disambiguate tenses and synonyms
  - Pattern.en for singular-plurals
  - N-gram overlap to differentiate “teddy bear” and “bear”, but if it’s connotation like…….


Does the bear love
WHAT???
made on imgur
rge bear a chairman?

Other tricky cases

- Synsets:
  - Sports ball includes football, basketball, baseball, etc.
  - Traffic lights→ Traffic signal
- English:
  - ‘Chicago Bears’
  - ‘Gummy bears’
  - These are false positives

How well can we detect from Qs?



- Mean recall 29.3%, mean precision 82.4%
- Larger objects get more attention(train +, baseball glove -, dining table meh)
- Objects detected area ~ 18.2%, failed to detect ~ 7.1%

Combining Visual classification with extracted information from Qs

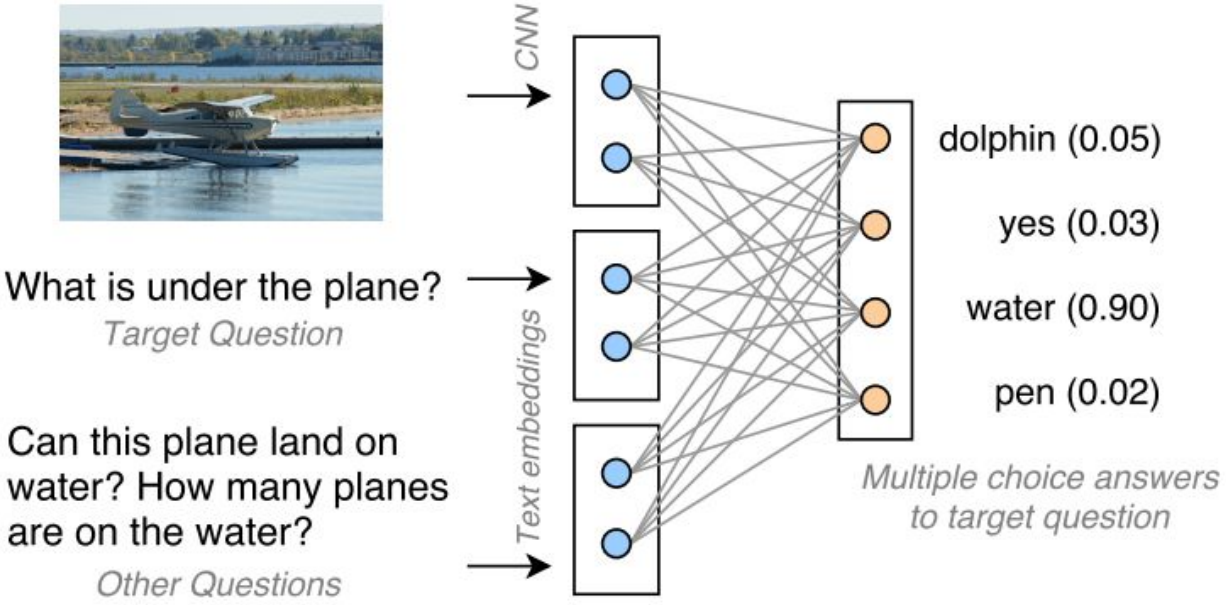
- Fine-tuned GoogLeNet on COCO, mAP 53.1% on validation set
- Combining 80-dim classifier prediction vector x_o with object class vector x_c extracted from 3 visual Qs using $max(x_o, x_c)$ with mAP 67.2%
- Visual questions even without answers provides informative image descriptions and object classification information

Inspiration

- Two key observations:
  - Different visual questions provide information complementary to each other
  - Visual questions can provide information about the scene that may be complementary to what can be extracted from the image
- Build a system having access to
  - Image information and the target question
  - Set of other questions that may have been asked about this image

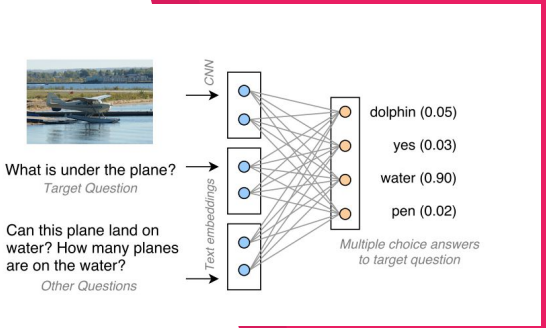
What's in the Question: **Structure**

iBOWIMG-2x



What's in the Question: **Structure**

- Image x_i associates with questions $\{q_{ij}\}_j$,corresponding answers $\{a_{ij}\}_j$ and unanswered questions $\{q'_{ik}\}_k$
- iBOWIMG feeds $(x_i, q_{ij}, a_{ij})\ \forall i, j$
- iBOWIMG2x feeds $(x_i, q_{ij}, E, a_{ij})\ \forall i, j, E \subseteq \mathcal{P}(Q_i^{all})$, $\mathcal{P}$ denotes the powerset of Q_i^{all} and defines the extra information provided to the model in the form of other asked questions (data augmentation)
- Target label is a_{ij} for the x_i, q_{ij}

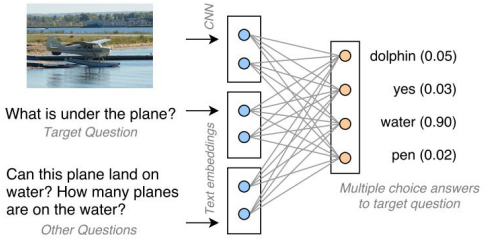


Two modifications

- Generate 2^n training exemplars incorporating all possible subsets of the n questions associated with the image (7.1% improvement)
- Can make use of only unanswered questions on novel images
  - What is to the left of the dog?
  - What is to the right of the person?

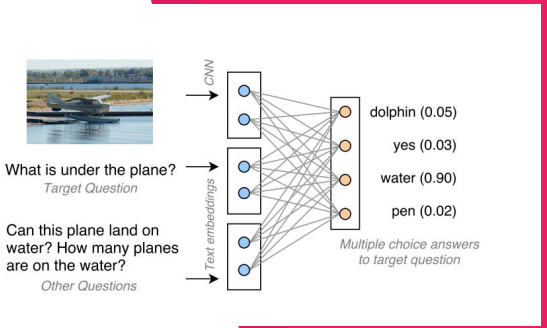
What's in the Question: **Structure**

- What if there is a large collection of images with only unanswered questions? No ground truth
  - We use a visual model trained to recognize the words that appear within the questions
  - ILSVRC-trained model is not best suited because it may not reflect the full spectrum of visual concepts or diverse visual scenes



Testing

- A novel image x and a single target question q
  - Pass zero-initialized vector for the extra features, reducing iBOWIMG-2x back to iBOWIMG, but still we trained them differently
- A novel image x and multiple target questions
  - We can make use of complementary information!



Experiments

- Dataset: COCO, each image has 3Qs, 3As
- Visual features and the two textual features are independently normalized to have L2 norm of 1, and we take a look at two different settings:
  - Every image has at least one answered question and optional unanswered questions
  - Some images have only unanswered questions associated with it
- Insights on how including extra questions significantly improves VQA accuracy

iBOWIMG-2x vs. iBOWIMG

What is on the back of the bike?

Helmet Life Vests

What color the shower curtain?

White 2

How many knives are in the knife holder?

3 6

At least one answered question

- Accuracy without augmentation:
  - $(x_i, q_i, [q'_{i1}\, q'_{i2}], a_i)$
- Accuracy with augmentation:
  - $(x_i, q_{ij}, E, a_{ij})\ \ \forall i, j, E \subseteq \mathcal{P}(Q_i^{all})$
- iBOWIMG upper left corner

Unanswered questions	Accuracy w/o aug	Accuracy
None	47.34	47.37
1 question	48.74	48.94
2 questions	49.19	**50.37**

Extra unanswered Qs vs. Augmentation

- Impact of having access to extra unanswered questions at training time
  - Using just one unanswered question achieves about half the improvement: 1.6% out of 3.1%
  - Adding more unanswered questions is likely to further improve accuracy

Unanswered questions	Accuracy w/o aug	Accuracy
None	47.34	47.37
1 question	48.74	48.94
2 questions	49.19	**50.37**

More about iBOWIMG2x

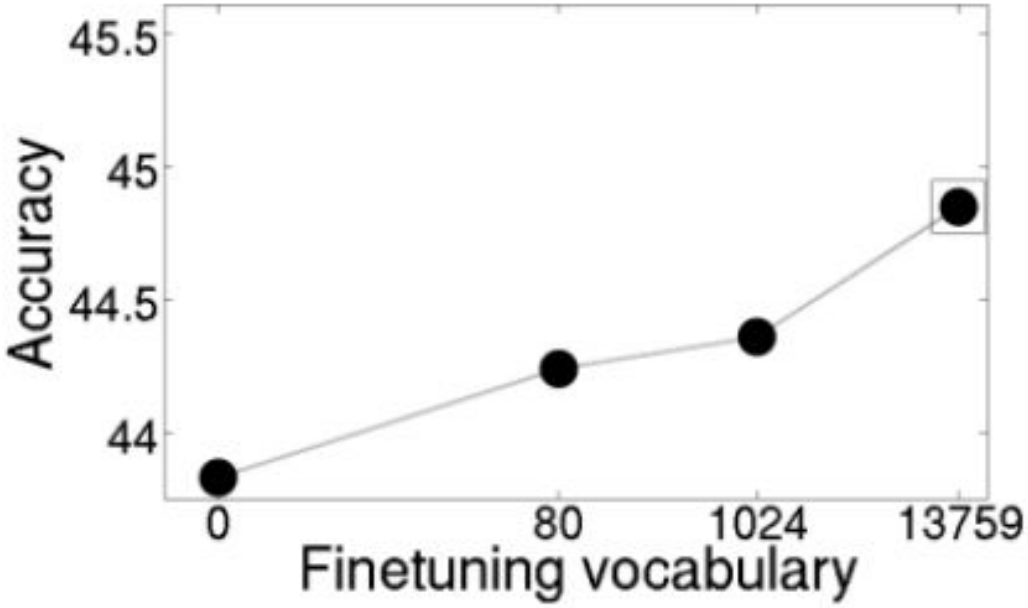
- Much of the benefit of iBOWIMG2x is in learning to make better use of the image features(text-only 46.7%->47.3%)
- iBOWIMG2x more likely to predict answers corresponding to actual words vs. number or y/n
- Richer representation better correlates the image appearance with the semantic textual features, making it more likely to predict a word answer

Answer Accuracy

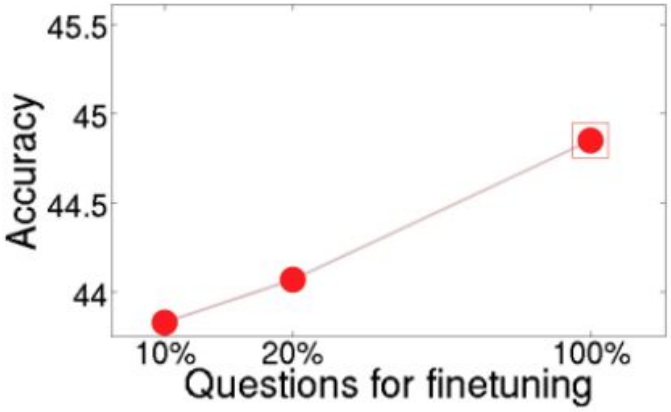
- 1.0% improvement on number Qs
- 3.0% on y/n Qs
- 3.9% on word-response Qs

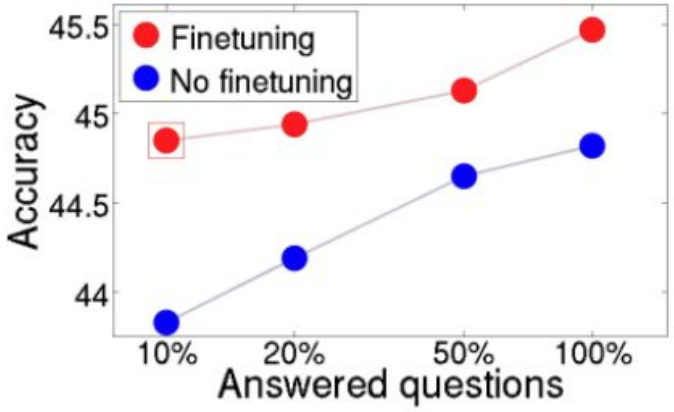
Model	Overall	Number	Yes/No	Word
iBOWIMG	45.87	26.85	74.53	34.07
iBOWIMG-2x	**50.37**	**27.92**	**77.54**	**37.98**

AlexNet fine-tuned on COCO questions



AlexNet fine-tuned on COCO questions





Full dataset with full data augmentation

Model Name	Overall	Other	Number	Yes/No
iBOWIMG	55.68	42.61	34.87	76.49
iBOWIMG-2x	**62.80**	**53.11**	**37.94**	**80.72**

Table 5: Multiple choice VQA accuracy on test-dev.

Reflections

Some simple but smart intuitions could really make a difference.

Never ignore the baselines.

The state of the art is still far from satisfaction to the real AI.

The end