Overcoming Priors in Visual Question Answering

Takeaways and Critique

Ishita Verma, Pranav Goel and Shlok Mishra

Computer Science at University of Maryland

Spotlight

"Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering" by Agrawal et al (2018).

Spotlight

"Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering" by Agrawal et al (2018).

- CVPR 2018.

Table of contents

Task and Motivation

Visual Question Answering (VQA)

What is the color of the refrigerator? white

Visual Question Answering (VQA)

What is the color of the refrigerator? white

How many people are standing? 2

Visual Question Answering (VQA)

Where is the conversation happening? kitchen

Visual Question Answering (VQA)

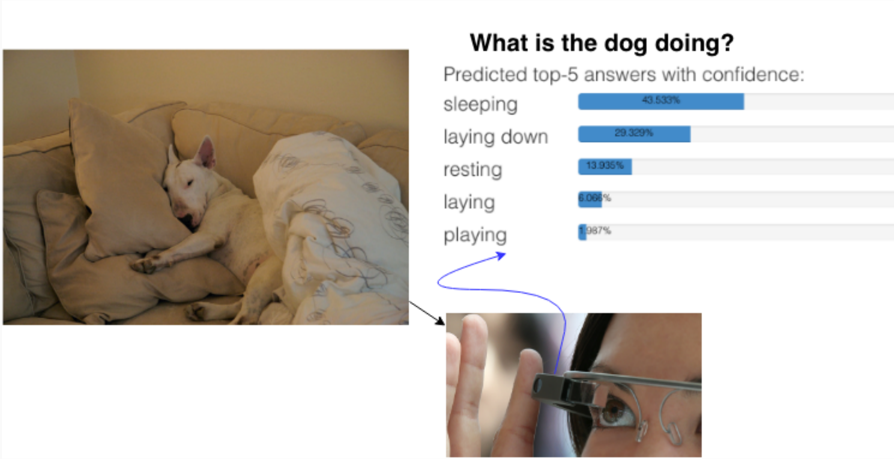
Where is the conversation happening? kitchen

Is the woman mad? yes

Why Should You Care?

Potential Real World Impact

- Aiding Visually Impaired Users

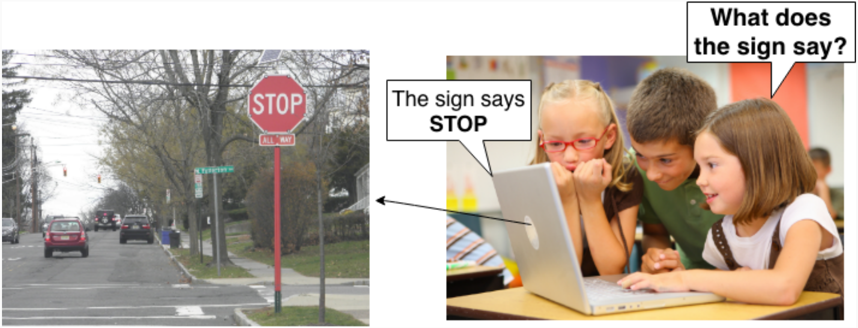


Potential Real World Impact

- Online Education

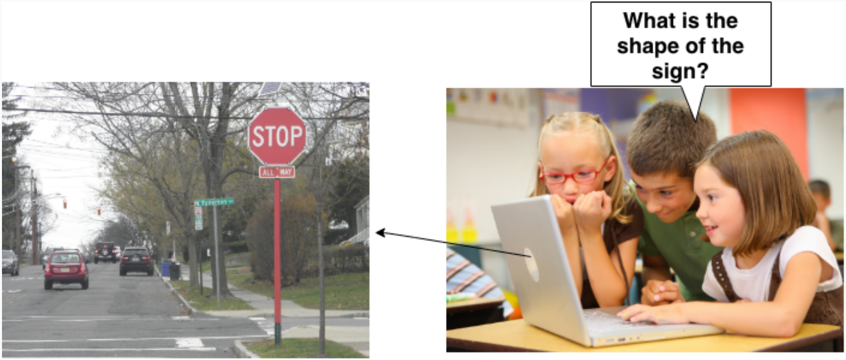
Potential Real World Impact

- Online Education



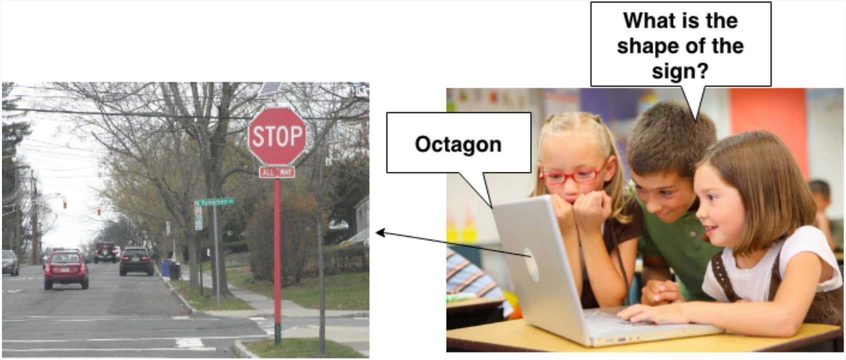
Potential Real World Impact

- Online Education



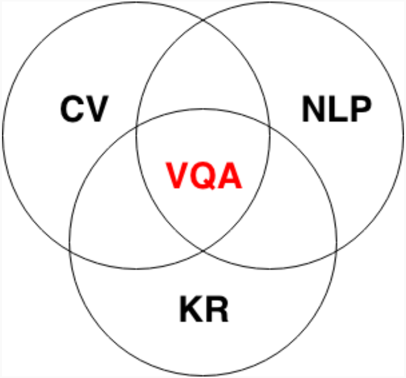
Potential Real World Impact

- Online Education



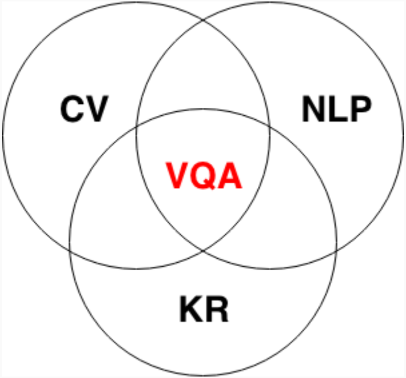
Scientific Appeal

Multi-modal knowledge required; beyond a single sub-domain!



Scientific Appeal

Effective combination is necessary!



Scientific Appeal

Need to understand the **question text** AND look at the **image**.

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

Scientific Appeal

Activity Recognition

Why are the men jumping? (to catch frisbee)

Scientific Appeal

Fine-grained recognition

What kind of cheese is there on the pizza?

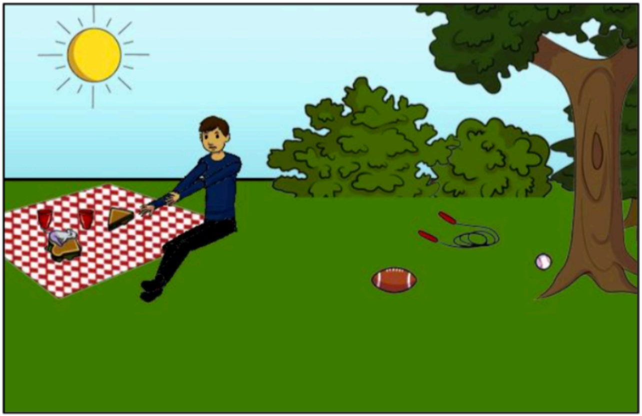
Scientific Appeal

Knowledge Base Reasoning

Is the pizza vegetarian?

Scientific Appeal

World Knowledge and Commonsense Reasoning



Is the person expecting company?

Context

Evolving Datasets

- MS COCO (Lin et al).

Evolving Datasets

- MS COCO (Lin et al).
- VQA 1.0 (Antol et al).

October 2015: Full release (v1.0)

Real Images

- 204,721 COCO images (all of current train/val/test)
- 614,163 questions
- 6,141,630 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

- 3 questions per image
- 10 answers per question

Evolving Datasets

- MS COCO (Lin et al).
- VQA **1.0** (Antol et al).

October 2015: Full release (v1.0)

Real Images
- 204,721 COCO images (all of current train/val/test)
- 614,163 questions
- 6,141,630 ground truth answers
- 1,842,489 plausible answers

- 3 questions per image
- 10 answers per question

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

- The evolution of datasets continued. Why?

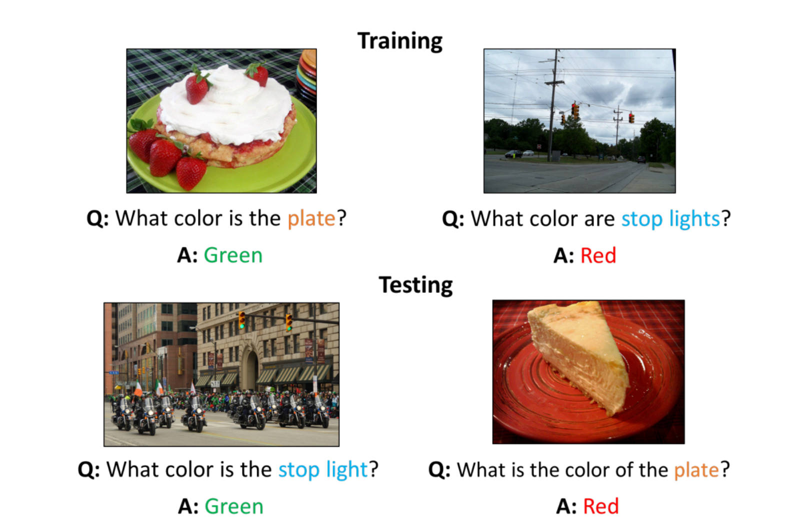
Problem: The Nuisance of Language Priors

Problem: The Nuisance of Language Priors

- VQA models can be heavily driven by superficial correlations in the training data and lack sufficient visual grounding.
- For e.g. - overwhelmingly replying to ‘how many X?’ questions with ‘2’ (irrespective of X), ‘what color is . . . ?’ with ‘white’, ‘is the . . . ?’ with ‘yes’.

VQA 2.0 (Goyal et al)

C-VQA (Compositional Split) (Agrawal et al)



The Work

A Problem Persists

- Even in C-VQA, **distribution of answers for each question type does not change much from train to test**. Models relying on priors can still perform well on test set!

A Problem Persists

- Even in C-VQA, **distribution of answers for each question type does not change much from train to test**. Models relying on priors can still perform well on test set!
- Why is this a problem? Benchmarking progress become difficult - what is the source of improvement (priors/visual grounding)?

A Problem Persists

- Even in C-VQA, **distribution of answers for each question type does not change much from train to test**. Models relying on priors can still perform well on test set!
- One Reason: IID train-test split of data having strong priors!

Proposed Solutions

Authors come up with two ways to address the problem -

Proposed Solutions

Authors come up with two ways to address the problem -

1. Change the train-test split! (Changing Priors Dataset).

Proposed Solutions

Authors come up with two ways to address the problem -

1. Change the train-test split! (Changing Priors Dataset).
2. A new model for grounded visual question answering (GVQA).

Changing Priors

The VQA-CP v1 and VQA-CP v2 splits are created such that the distribution of answers per question type ('how many', 'what color is', etc.) is *different* in the test data compared to the training data.

Changing Priors

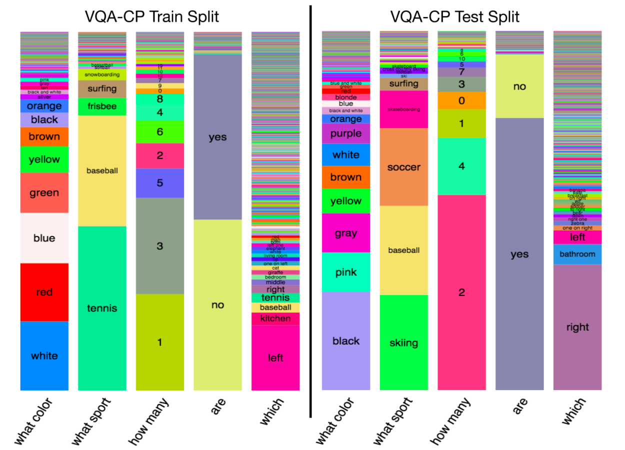
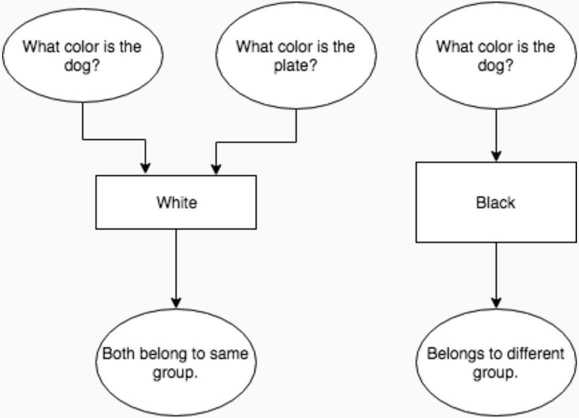
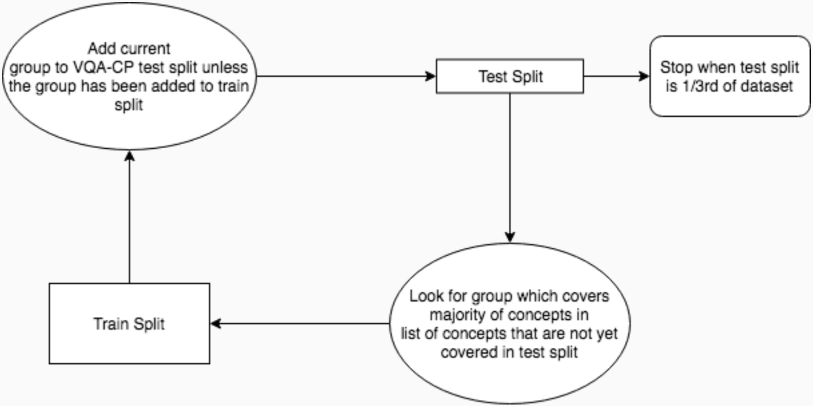


Figure 2: Distribution of answers per question type vary significantly between VQA-CP v1 train (left) and test (right) splits. For instance, *'white'* and *'red'* are commonly seen answers in train for *'What color'*, where as *'black'* is the most frequent answer in test. These have been computed for a random sample of 60K questions.

Question Grouping:

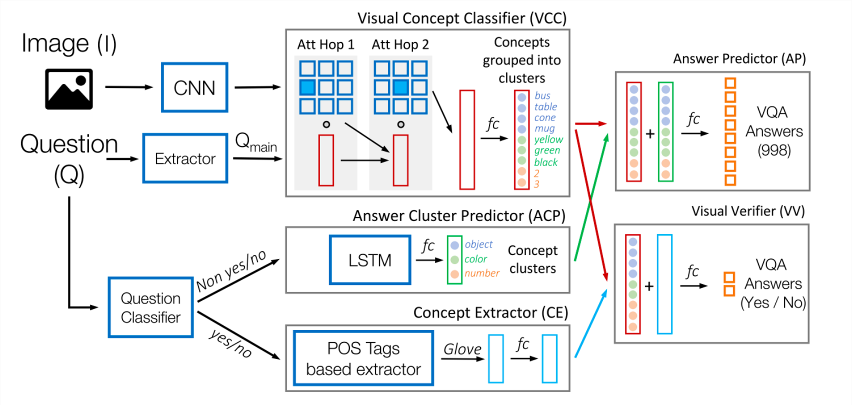


Greedy Re-splitting:

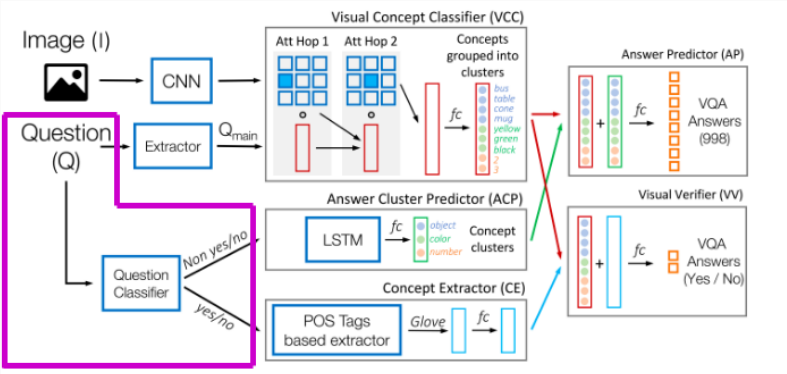


The GVQA Model

The proposed Grounded Visual Question Answering model.

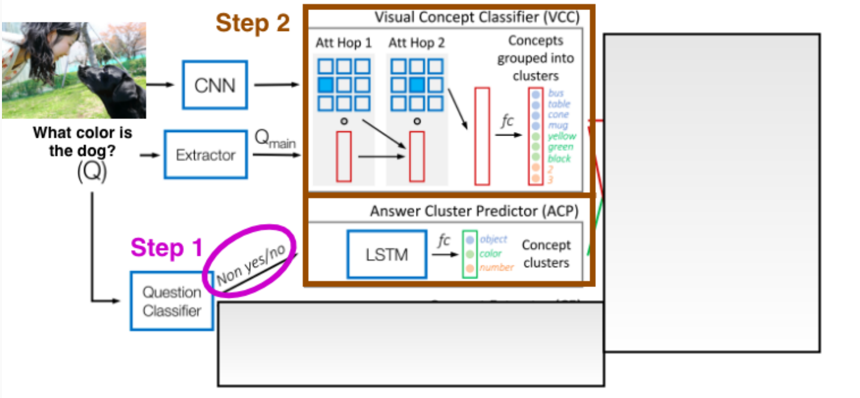


Step 1 (for any question): Question Classification



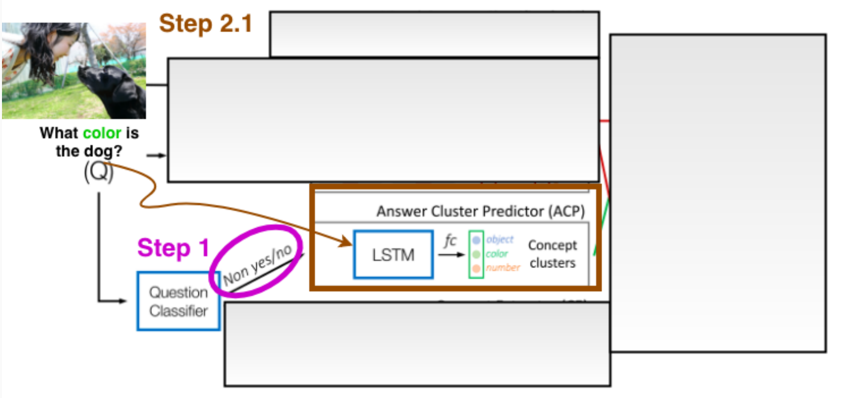
Non yes/no

Step 2: VCC and ACP activated.



Non yes/no

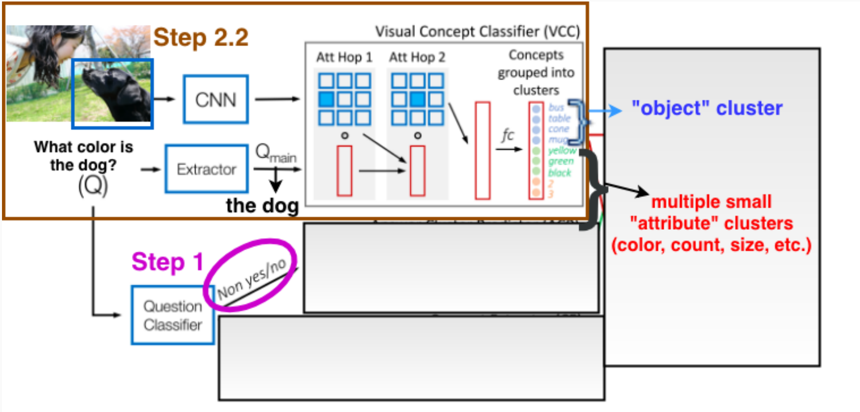
Step 2.1: Answer Cluster Prediction (ACP)



The GVQA Model

Non yes/no

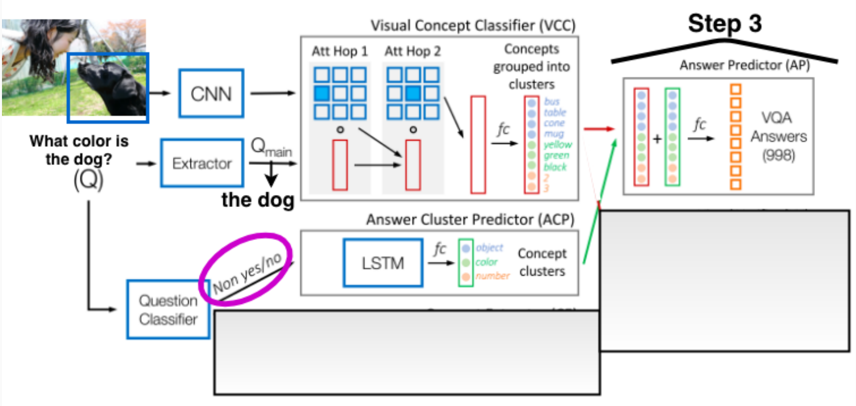
Step 2.2: Visual Concept Classifier (VCC)



The GVQA Model

Non yes/no

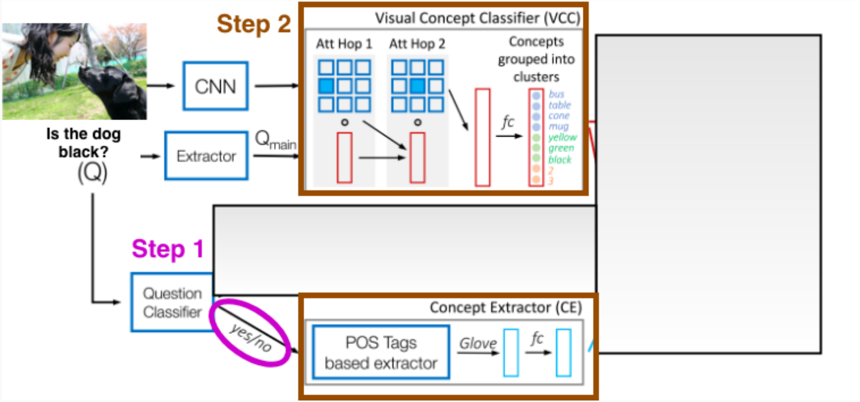
Step 3: Answer Predictor (AP)



The GVQA Model

yes/no

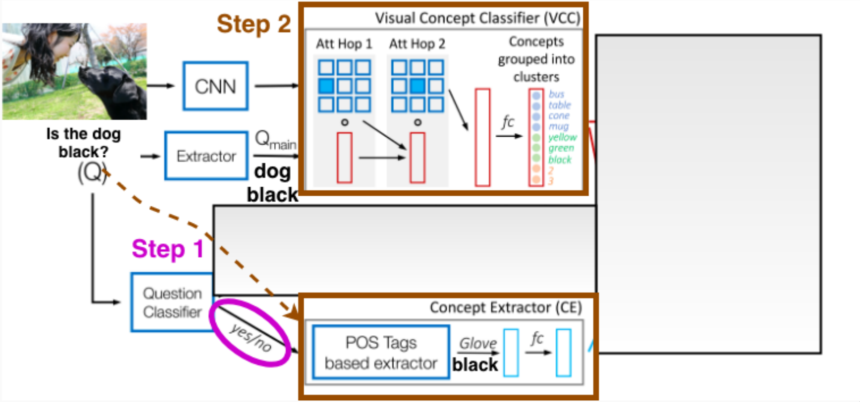
VCC and CE activated.



yes/no

VCC - Visual concepts to verify over.

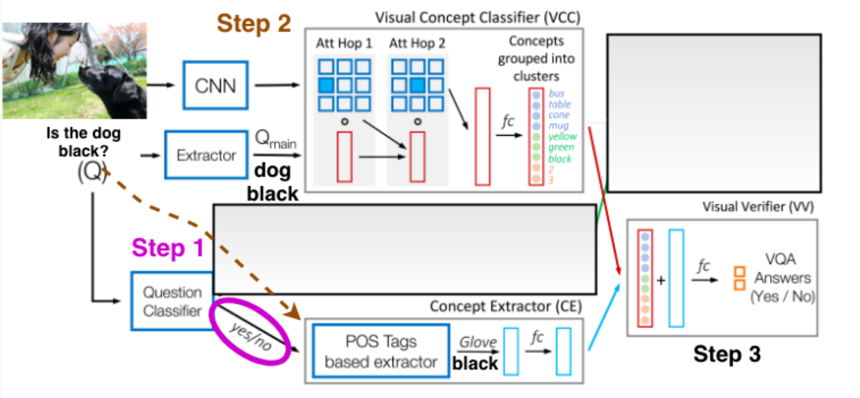
CE - Extract concepts to verify.



The GVQA Model

yes/no

Visual Verification



Dataset	Model	Overall	Yes/No	Number	Other
VQA-CP v1	SAN [39]	26.88	35.34	11.34	24.70
	GVQA (Ours)	**39.23**	**64.72**	**11.87**	**24.86**
VQA-CP v2	SAN [39]	24.96	38.35	11.14	21.74
	GVQA (Ours)	**31.30**	**57.99**	**13.68**	**22.14**

Results and Discussion

Model	Overall	Yes/No	Number	Other
GVQA - $Q_{main} + Q_{full}$	33.55	51.64	11.51	24.43
GVQA - CE + LSTM	27.28	35.96	11.88	24.85
GVQA - ACP + LSTM	39.40	64.72	11.73	25.33
GVQA - VCC_{loss}	40.95	65.50	12.32	28.05
GVQA	39.23	64.72	11.87	24.86

Results and Discussion

Model	Overall	Yes/No	Number	Other
GVQA - $Q_{main} + Q_{full}$	33.55	51.64	11.51	24.43
GVQA - CE + LSTM	27.28	35.96	11.88	24.85
GVQA - ACP + LSTM	39.40	64.72	11.73	25.33
GVQA - VCC_{loss}	40.95	65.50	12.32	28.05
GVQA	39.23	64.72	11.87	24.86

Results and Discussion

Model	Overall	Yes/No	Number	Other
GVQA - $Q_{main} + Q_{full}$	33.55	51.64	11.51	24.43
GVQA - CE + LSTM	27.28	35.96	11.88	24.85
GVQA - ACP + LSTM	39.40	64.72	11.73	25.33
GVQA - VCC_{loss}	40.95	65.50	12.32	28.05
GVQA	39.23	64.72	11.87	24.86

Results and Discussion

Model	Overall	Yes/No	Number	Other
GVQA - $Q_{main} + Q_{full}$	33.55	51.64	11.51	24.43
GVQA - CE + LSTM	27.28	35.96	11.88	24.85
GVQA - ACP + LSTM	39.40	64.72	11.73	25.33
GVQA - VCC$_{loss}$	40.95	65.50	12.32	28.05
GVQA	39.23	64.72	11.87	24.86

Results and Discussion

Model	VQA v1	VQA v2
SAN	55.86	52.02
GVQA	51.12	48.24

Model	VQA v1	VQA v2
SAN	55.86	52.02
GVQA	51.12	48.24
Ensemble (SAN, SAN)	56.56	52.45
Ensemble (GVQA, SAN)	56.91	52.96

Results and Discussion

Model	VQA v1	VQA v2
SAN	55.86	52.02
GVQA	51.12	48.24
Ensemble (SAN, SAN)	56.56	52.45
Ensemble (GVQA, SAN)	56.91	52.96
Oracle (SAN, SAN)	60.85	56.68
Oracle (GVQA, SAN)	63.77	61.96

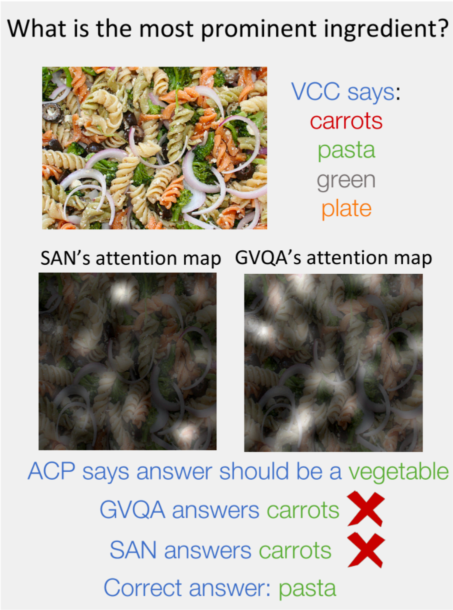
Increased Transparency

GVQA - (VCC + ACP/CE) structure allows us to speculate why an answer was given!

SAN - Stacked Attention Network - why does it predict what it predicts?

Increased Transparency

Increased Transparency



Critical Assessment

Designing the Train-Test Split

- Explicit designing of the train-test split.

Designing the Train-Test Split

- Explicit designing of the train-test split.
- Addresses the problem of IID split in presence of strong priors.

Designing the Train-Test Split

- Explicit designing of the train-test split.
- Addresses the problem of IID split in presence of strong priors.
- A useful idea beyond this particular work?

Designing the Train-Test Split

- Different ‘split designs’ to control different priors?

Designing the Train-Test Split

- Different ‘split designs’ to control different priors?
- How do different models fare on the different splits? Can we gain more insights via that?

Designing the Train-Test Split

- Different 'split designs' to control different priors?
- How do different models fare on the different splits? Can we gain more insights via that?
- Building models that perform well across all these splits.

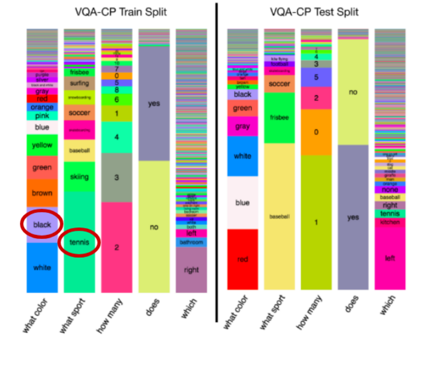
Does GVQA Overcome Priors?

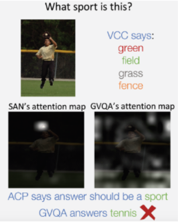
- The problem, as stated in the paper - “It seems that when faced with a difficult learning problem, models typically resort to latching onto the language priors in the training data”.

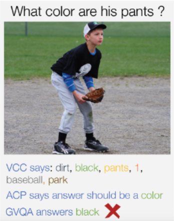
Does GVQA Overcome Priors?

- The problem, as stated in the paper - “It seems that when faced with a difficult learning problem, models typically resort to latching onto the language priors in the training data”.
- What happens when the GVQA model ‘faces difficulty’ and makes an incorrect prediction? No real info provided.

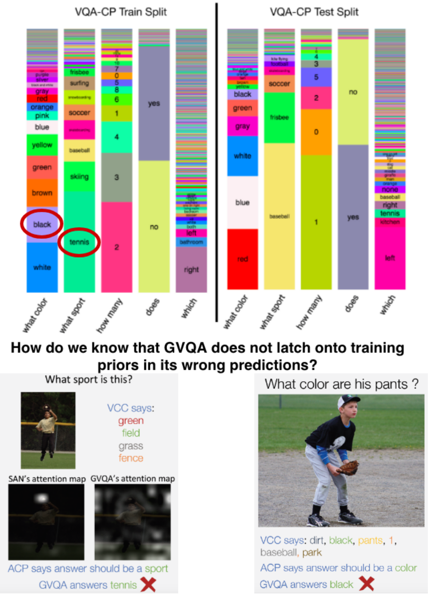
Does GVQA Overcome Priors?







Does GVQA Overcome Priors?



Does GVQA Overcome Priors?

Could report % of *incorrect answers* that come from training priors.

Are All Priors Bad?

- Some priors could help with world knowledge - sky is blue, a person usually has one nose, etc.

Are All Priors Bad?

- Some priors could help with world knowledge - sky is blue, a person usually has one nose, etc.
- Interesting future direction (pointed out by the authors) - **“models that can utilize the best of both worlds (visual grounding and priors)”**.

Are All Priors Bad?

- Some priors could help with world knowledge - sky is blue, a person usually has one nose, etc.
- Interesting future direction (pointed out by the authors) - **"models that can utilize the best of both worlds (visual grounding and priors)"**.
- Priors need to be *derived from world knowledge* and must NOT be *artifacts of a particular dataset.*

Thank you!

The Problem of Language Priors

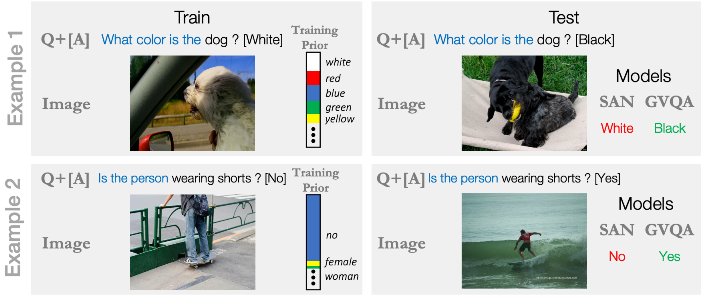


Figure 1: Existing VQA models, such as SAN [39], tend to largely rely on strong language priors in train sets, such as, the prior answer (*'white'*, *'no'*) given the question type (*'what color is the'*, *'is the person'*). Hence, they suffer significant performance degradation on test image-question pairs whose answers (*'black'*, *'yes'*) are not amongst the majority answers in train. We propose a novel model (GVQA), built off of SAN that explicitly grounds visual concepts in images, and consequently significantly outperforms SAN in a setting with mismatched priors between train and test.

Previous Models on VQA-CP

Model	Dataset	Overall	Yes/No	Number	Other	Dataset	Overall	Yes/No	Number	Other
per Q-type prior [5]	VQA v1	35.13	71.31	31.93	08.86	VQA v2	32.06	64.42	26.95	08.76
	VQA-CP v1	08.39	14.70	08.34	02.14	VQA-CP v2	08.76	19.36	11.70	02.39
d-LSTM Q [5]	VQA v1	48.23	79.05	33.70	28.81	VQA v2	43.01	67.95	30.97	27.20
	VQA-CP v1	20.16	35.72	11.07	08.34	VQA-CP v2	15.95	35.09	11.63	07.11
d-LSTM Q + norm I [24]	VQA v1	54.40	79.82	33.87	40.54	VQA v2	51.61	73.06	34.41	39.85
	VQA-CP v1	23.51	34.53	11.40	17.42	VQA-CP v2	19.73	34.25	11.39	14.41
NMN [3]	VQA v1	54.83	80.39	33.45	41.07	VQA v2	51.62	73.38	33.23	39.93
	VQA-CP v1	29.64	38.85	11.23	27.88	VQA-CP v2	27.47	38.94	11.92	25.72
SAN [39]	VQA v1	55.86	78.54	33.46	44.51	VQA v2	52.02	68.89	34.55	43.80
	VQA-CP v1	26.88	35.34	11.34	24.70	VQA-CP v2	24.96	38.35	11.14	21.74
MCB [11]	VQA v1	60.97	81.62	34.56	52.16	VQA v2	59.71	77.91	37.47	51.76
	VQA-CP v1	34.39	37.96	11.80	39.90	VQA-CP v2	36.33	41.01	11.96	40.57

Table 1: We compare the performance of existing VQA models on VQA-CP test splits (when trained on respective VQA-CP train splits) to their performance on VQA val splits (when trained on respective VQA train splits). We find that the performance of all tested existing models degrades significantly in the new Changing Priors setting compared to the original VQA setting.

Qualitative Examples (GVQA)

Figure 4: Qualitative examples from GVQA. **Left:** We show top three answer cluster predictions (along with random concepts from each cluster) by ACP. Corresponding to each cluster predicted by ACP, we show the top visual concept predicted by VCC. Given these ACP and VCC predictions, the Answer Predictor (AP) predicts the correct answer *‘baseball’*. **Right:** Smiling is the concept extracted by the CE whose visual presence in VCC’s predictions is verified by the Visual Verifier, resulting in *‘yes’* as the final answer.

Increased Transparency

Figure 5: **Left**: GVQA's prediction (*'green'*) can be explained as follows – ACP predicts that the answer should be a *color*. Of the various visual concepts predicted by VCC, the only concept that is about color is *green*. Hence, GVQA's output is *'green'*. SAN incorrectly predicts *'yellow'*. SAN's architecture doesn't facilitate producing an explanation of why it predicted what it predicted, unlike GVQA. **Right**: Both GVQA and SAN incorrectly answer the question. GVQA is incorrect perhaps because VCC predicts *'black'*, instead of *'gray'*. In order to dig further into why VCC's prediction is incorrect, we can look at the attention map (in Fig. 8), which shows that the attention is on the pants for the person's left leg, but on the socks (black in color) for the person's right leg. So, perhaps, VCC's "black" prediction is based on the attention on the person's right leg.

Increased Transparency

GVQA is “looking” at the smartphone (unlike SAN), but yet incorrectly answers ‘no’, because the VCC does not recognize the phone as a smartphone. It however correctly predicts ‘phone’, ‘electronic’, ‘black’ and ‘right’.

Full Results

Model	Overall	Yes/No	Number	Other
GVQA - Q_{main} + Q_{full}	33.55	51.64	11.51	24.43
GVQA - CE + LSTM	27.28	35.96	11.88	24.85
GVQA - ACP + LSTM	39.40	64.72	11.73	25.33
GVQA - VCC_{loss}	40.95	65.50	12.32	28.05
GVQA - VCC_{loss} - ACP + LSTM	38.86	65.73	11.58	23.11
GVQA	39.23	64.72	11.87	24.86

Table 3: Experimental results when each component in GVQA (denoted by "- <component>") is replaced with its corresponding traditional counterpart (denoted by " + <traditional counterpart>").

Additional Splits of VQA-CP v2

Model	Split1	Split2	Split3	Split4	Average
SAN	24.96	26.07	22.69	24.19	24.48
GVQA	31.30	32.40	33.78	28.99	31.62
	+6.34	+6.33	+11.09	+4.80	+7.14

Figure 7: Performance of SAN and GVQA for different VQA-CP v2 splits. GVQA consistently outperforms SAN across all splits.

As mentioned in Section 6.1, to check if our particular VQA-CP split was causing some irregularities in performance, we created three additional sets of VQA-CP v2 splits with different random seeds. We evaluated both SAN and GVQA on all four splits (please see Fig. 7). We can see that GVQA consistently outperforms SAN across all four splits with average improvement being 7.14% (standard error: 1.36).

Performance of SAN with Q_{main}

Model	Overall	Yes/No	Number	Other
SAN [39]	24.96	38.35	11.14	21.74
SAN - Q_{full} + Q_{main}	26.32	44.73	09.46	21.29
GVQA (Ours)	**31.30**	**57.99**	**13.68**	**22.14**

Table 5: Performance of SAN - Q_{full} + Q_{main} compared to SAN and GVQA (our model) on VQA-CP v2 dataset. GVQA outperforms both SAN and SAN - Q_{full} + Q_{main}.

As mentioned in Section 6.2, as an additional check, we trained a version of SAN where the input is Q_{main} instead of Q_{full}. Table 5 shows the results of this version of SAN (SAN - Q_{full} + Q_{main}) along with those of SAN and GVQA on VQA-CP v2. We can see that this version of SAN performs 1.36% (overall) better than the original SAN, however still 4.98% (overall) worse than GVQA (with Q_{main}).

Performance of $GVQA - VCC_{loss}$ on VQA v1 and VQA v2

Model	VQA v1				VQA v2			
	Overall	Yes/No	Number	Other	Overall	Yes/No	Number	Other
SAN	55.86	78.54	33.46	44.51	52.02	68.89	34.55	43.80
GVQA - VCC_{loss}	48.51	65.59	32.67	39.71	48.34	66.38	31.61	39.05
GVQA	51.12	76.90	32.79	36.43	48.24	72.03	31.17	34.65
Ensemble (SAN, SAN)	56.56	79.03	34.05	45.39	52.45	69.17	34.78	44.41
Ensemble ((GVQA - VCC_{loss}), SAN)	56.44	78.27	34.45	45.62	51.79	68.59	34.44	43.61
Ensemble (GVQA, SAN)	56.91	80.42	34.40	44.96	52.96	72.72	34.19	42.90
Oracle (SAN, SAN)	60.85	83.92	39.43	48.96	56.68	74.37	40.08	47.61
Oracle ((GVQA - VCC_{loss}), SAN)	64.47	90.17	42.92	50.64	61.93	85.13	43.51	49.16
Oracle (GVQA, SAN)	63.77	88.98	43.37	50.03	61.96	85.65	43.76	48.75

Table 6: Results of GVQA, GVQA - VCC_{loss} and SAN on VQA v1 and VQA v2 when trained on the corresponding train splits. Plea