Automatic Generation of Contrast Sets from Scene Graphs: Probing the Compositional Consistency of GQA

The Hebrew University Of Jerusalem
Ben-Gurion University of the Negev

Yonatan Bitton, Gabriel Stanovsky, Roy Schwartz, Michael Elhadad

Overview

Models often exploit data artifacts to achieve good test scores.

https://thegradient.pub/shortcuts-neural-networks-love-to-cheat/

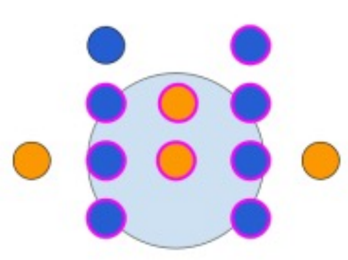
McCoy, R. Thomas, et al. "Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference.", ACL 2019.

Gururangan, Suchin, et al. "Annotation artifacts in natural language inference data." , NAACL 2018.

Jia, Robin, et al. "Adversarial examples for evaluating reading comprehension systems." , EMNLP 2017.

Contrast sets quantify this phenomenon. Used as a more accurate evaluation the for models true capabilities 🔍.

Contrast sets
Gardner, Matt, et al. "Evaluating models' local decision boundaries via contrast sets", Findings of EMNLP 2020



In many cases, contrast sets have been built manually, requiring extensive human effort and expertise 🕵.

Original Instance	**Contrastive Instance** (color = edit)
Hardly one to be faulted for his ambition or his vision, it is genuinely unexpected, then, to see all Park's effort add up to so very little. . . . The premise is promising, gags are copious and offbeat humour abounds but it all fails miserably to create any meaningful connection with the audience. *(Label: Negative)*	Hardly one to be faulted for his ambition or his vision, **here we see all Park's effort come to fruition.** . . . The premise is **perfect**, gags are **hilarious** and offbeat humour abounds, **and it creates a deep** connection with the audience. *(Label: Positive)*

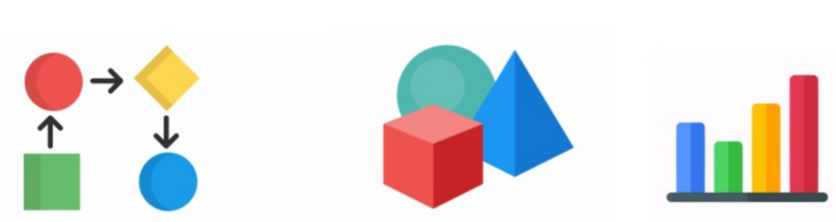
We propose a method for automatic construction of large contrast sets for the Visual Question Answering task, by leveraging scene-graphs input representations.

VQA Visual Question Answering

We demonstrate the effectiveness of our method on the GQA dataset.

GQA

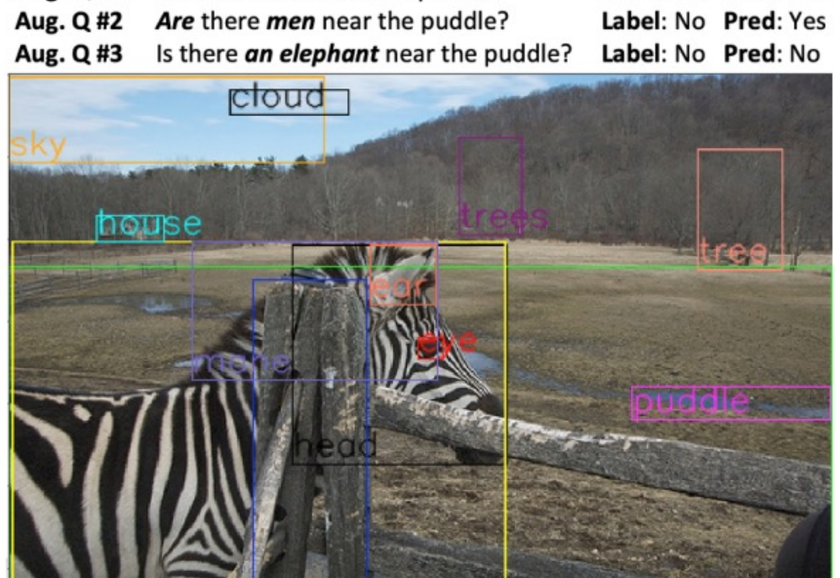
Hudson, Drew A, et al. "GQA: A new dataset for real-world visual reasoning and compositional question answering." *CVPR* 2019.

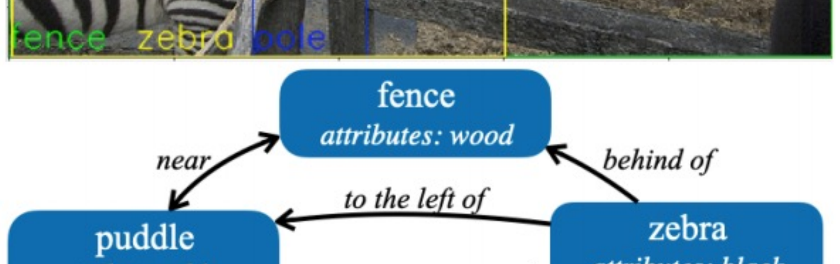


SEMANTIC REPRESENTATIONS COMPOSITIONAL BALANCED

Starting from $(image, scene\ graph, Q, A)$ we generate a set of variants $\{(image, scene\ graph, Q_i', A_i')\}$ s.t Q_i' is a minimal change of Q, and $A \neq A_i'$.

Original Q	Is there ***a fence*** near the puddle?	**Label**: Yes	**Pred**: Yes
Aug. Q #1	Is there ***a wall*** near the puddle?	**Label**: No	**Pred**: Yes
Aug. Q #2	***Are*** there ***men*** near the puddle?	**Label**: No	**Pred**: Yes
Aug. Q #3	Is there ***an elephant*** near the puddle?	**Label**: No	**Pred**: No





Automatic Contrast Set Construction

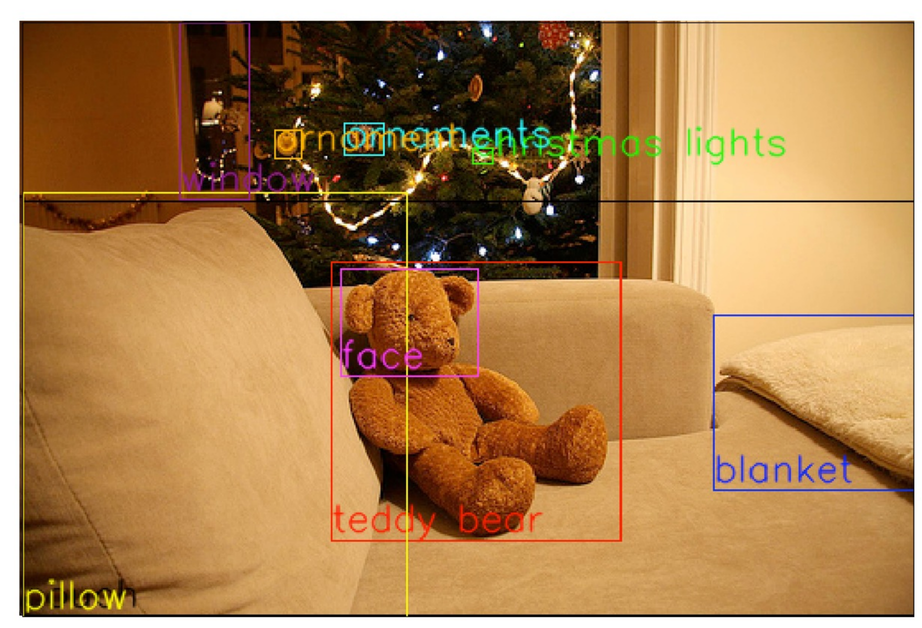
Identifying Recurring Patterns in GQA

Question template	Tested attributes	Example
On which side is the X?	Relational (left vs. right)	On which side is the ***dishwasher***? → On which side are the ***dishes***?
What color is the X?	Color identification	What color is the ***cat***?→ What color is the ***jacket***?
Do you see X or Y?	Compositions	Do you see ***laptops*** or cameras?→ Do you see ***headphones*** or cameras?
Are there X near the Y? Is the X ***Rel*** the Y? Is the X *Rel* the Y?	Spatial, relational	*Are* there any ***cats*** near the boat? → *Is* there any ***bush*** near the boat? Is the boy to the ***right*** of the man? → Is the boy to the ***left*** of the man? Is the boy to the right of the ***man***? → Is the boy to the right of the ***zebra***?

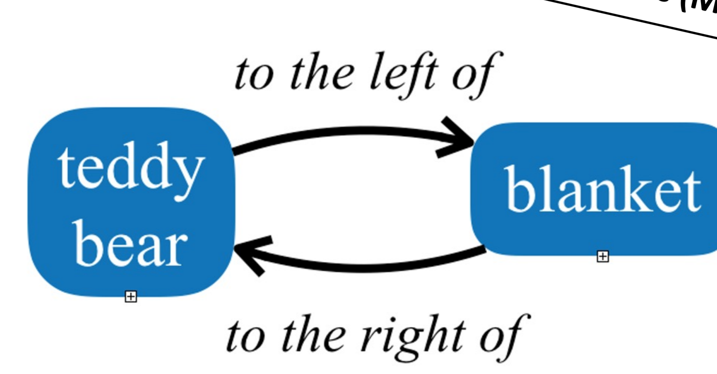
Illustrating the perturbation process

Is the *teddy bear* to the *left* of a *suitcase*? No → Is the *teddy bear* to the *left* of a ***blanket***? Yes

Is the X *Rel* the Y?



We verify perturbation correctness with human annotations (Mechanical Turk)



Main Findings

Models struggle with our contrast sets

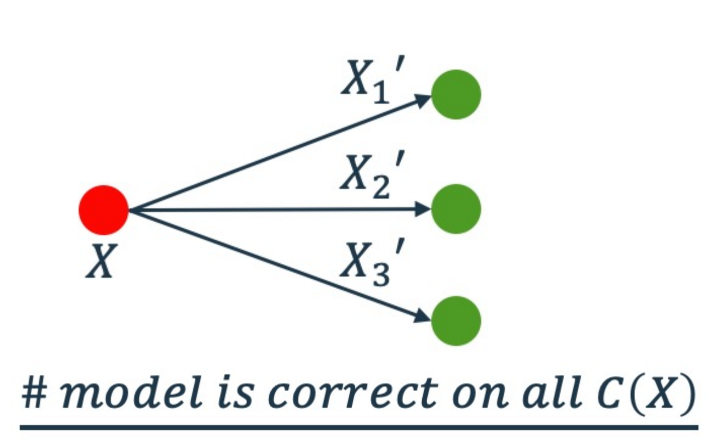
	MAC		LXMERT	
	Original	Aug.	Original	Aug.
On which side is the X?	68%	57%	94%	81%
What color is the X?	49%	49%	69%	62%
Are there X near the Y?	85%	66%	98%	79%
Do you see X or Y?	88%	53%	95%	65%
Is the X ***Rel*** the Y?	85%	44%	96%	69%
Is the X ***Rel*** the Y?	71%	38%	93%	55%
Overall	**65%**	**52%**	**84%**	**67%**

Training on perturbed set leads to more robust models

Model	Training set	Original	Augmented
MAC	Baseline	64.9%	51.5%
	Augmented	64.4%	**68.4%**
LXMERT	Baseline	83.9%	67.2%
	Augmented	82.6%	**77.2%**

Since our method is automatic, we can augment the training set as well

Consistency drops as the number of augmentations grow



Augmentations per instance	Contrast sets	Acc.	Consistency
1	11,263	66%	63.4%
3	23,236	67%	51.1%
5	28,968	67%	46.1%