Discovering the Unknown Knowns: Turning Implicit Knowledge in the Dataset into Explicit Training Examples for Visual Question Answering

Jihyung Kil, Cheng Zhang, Dong Xuan, Wei-Lun Chao
The Ohio State University

EMNLP 2021 Online & in the Dominican Republic — Conference on Empirical Methods in Natural Language Processing

Highlights

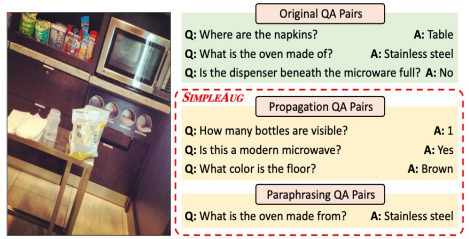
- Propose ***SimpleAug***, a simple and model agnostic data augmentation method that turns information already in the datasets into explicit IQA triplets for training VQA models
- ***SimpleAug*** can notably improve VQA models' accuracy on both VQA v2 and VQA-CP
- Conduct comprehensive analyses on ***SimpleAug***, including its applicability to the unlabeled images

Introduction

- VQA Challenges
  - VQA models trained on the human labeled data overfits the language bias or struggle in capturing the diversity of human language
  - We argue that they may result from a fundamental issue: Not enough training examples
  - Evidence: If we ask more questions about the training images (e.g., by borrowing relevant questions from other training images), VQA models fail drastically
  - This implies VQA model hasn't still learned enough information from the human labeled data even if they have already seen these images and questions

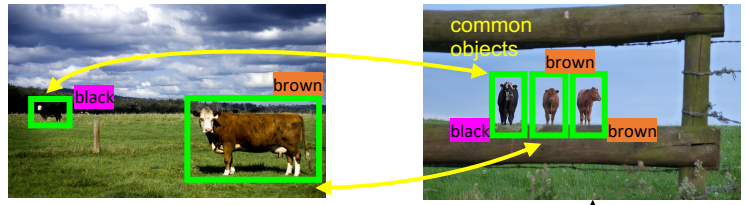
SimpleAug

- Data augmentation turning implicit information already in the dataset into explicit training examples



Implicit knowledge in the VQA dataset

i. IQA Triplets ii. Mid-level Semantic Annotations (MSCOCO) iii. Pre-trained Object Detector (Faster-RCNN)



Q: What is the color of left cow? A: black

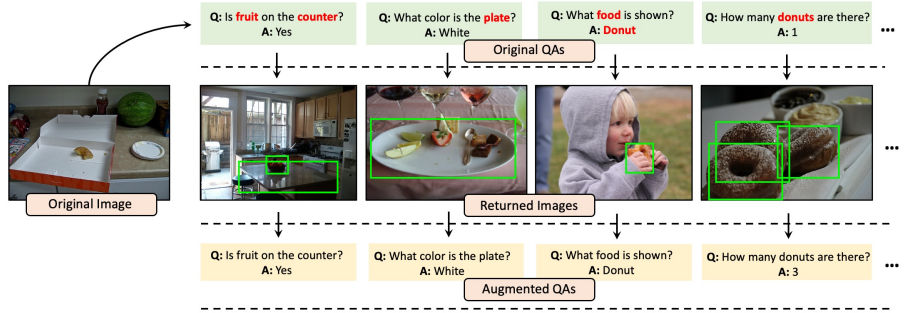
Q: What animal is this? A: cow

Q: How many animals are in the picture? A: 2

propagate

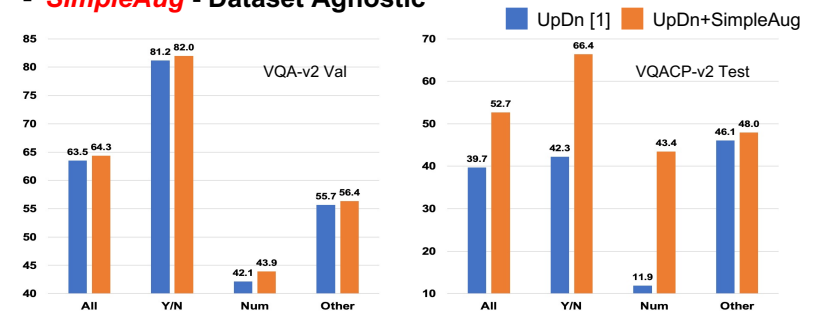
SimpleAug Pipeline

- Questions annotated for one image can be valuable add-ons to other relevant images
- "Propagate" questions from one image to other relevant images using three sources of implicit knowledge (i.e., i., ii., and iii.)

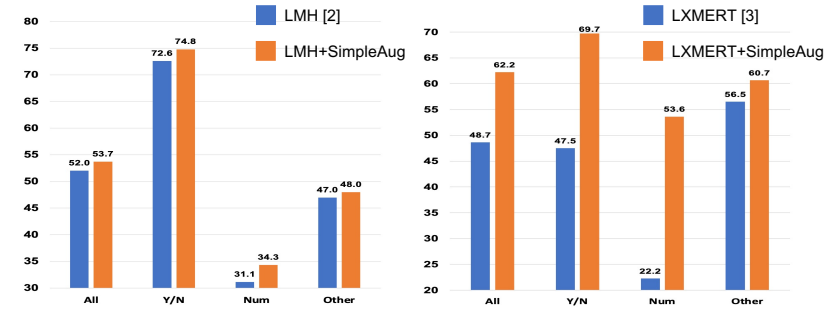


Experiments

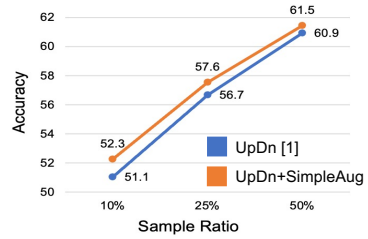
SimpleAug - Dataset Agnostic



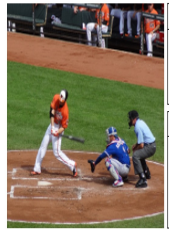
SimpleAug - Model Agnostic



Learning on unlabeled images



Qualitative Results



Original Question	Answer	
What color are the empty seats?	Green	✗
How many people are on the field?	3	✓
What team is playing?	Orioles	✓

Augmented Question	Answer	
How many baseball bats are in the picture?	1	✗
How many baseball gloves are showing?	1	✗
What color is the helmet?	Blue	✗
How many people are in the field?	3	✓

[1] Bottom-up and top-down attention for image captioning and visual question answering. In CVPR.
[2] Don't take the easy way out: Ensemble based methods for avoiding known dataset. In EMNLP.
[3] LXMERT: Learning cross-modality encoder representations from transformers. In EMNLP.