COMPUTER VISION FOR VISUALLY IMPAIRED PEOPLE : ANALYSIS ON THE VIZWIZ DATASET

A Project Report

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

MSc.

by

Kuan-Wen Wang, Yezhou Ma

May 2020

ABSTRACT

In this study, we explore analyze the VizWiz dataset in a computer vision perspective, and identify challenges and potential solutions toward building computer vision model for visually impaired people.

We conducted analysis on the images and question-answer(QA) pairs on the VizWiz[2] dataset to identify common domains of problems. We also targeted object detection as a first step. By analyzing and mapping the QA pairs to ImageNet[3] labels, we found that building a new set of labels specifically designed for this domain would be crucial. We then inspect and build a vocabulary set for the object detection task.

ACKNOWLEDGEMENTS

This study was supported by Cornell Tech and Professor Shiri Azenkot's research group. We would like to express the deepest appreciation to our advisors, Professor Shiri Azenkot, Yuhang Zhao, and Yin Cui, who introduced us to the research area of AI-powered Systems for Access and Equity. Their insightful guidance and persistent help throughout the study have done a great favor to both our specialized project and personal academic achievements.

We would also like to thank the staff at Cornell Tech to coordinate the specialization project, connecting the advisors and resources to us.

CHAPTER 1
INTRODUCTION

Visual recognition models that are trained and tested on photos taken by sighted people already have great success. With modern deep convolutional neural networks (CNNs), we can achieve high accuracy on several datasets such as ImageNet and COCO [6]. However, these datasets and models have an implicit assumption that the photos are taken by sighted people. In the case of photos taken by visually impaired people, the photo is dramatically different. These photos might lose focus, having partially blocked by hands or other objects, or only contains part of the target object. All these features prevent the current state-of-the-art visual recognition model from accurately detecting objects.

In this study, we look at VizWiz, a very goal-oriented Visual Question Answering (VQA) dataset aiming to assist blind people. Their analysis showed that this dataset is statistically very different from existing VQA datasets, and is challenging as current state of the art models performs poorly on this dataset. We dive deeper into the dataset, analyze the question answering pairs by Part-of-Speech Tagging, identifying nouns that could be present in the photo. Our attempt to map the nouns into existing image classification dataset such as ImageNet is proven to be not successful. We proceed to build our own multi-label classification label that are within this domain by manually inspecting top frequently mentioned nouns, and confirm the associated photo set is correct and unambiguous.

A note about this report: it is the final report for the Specialization Project, required for the Connective Media Master program. This project was a two-person research *project done under the advice of a faculty member at Cornell Tech.*

CHAPTER 2
RELATED WORK

There are numerous previous works on the topic of pictures taken by blind people. One of the biggest issue in the VizWiz dataset is that pictures taken by blind people often are of poor quality, including poor lighting, focus and framing, making it hard or even impossible to answer the questions.

Many works has provided effective methods to assist blind people to take better pictures. These methods includes voice feedback mechanism like EasySnap [5] and BlindCamera[1], or re-designing photography framework [9]. However, the evaluations of all these works only based on very few participants (20), and is limited by the assumption that users aim the device roughly in the right direction at first so that further adjustment is possible.

Also, a very distinct feature of the VizWiz dataset is that there are a lot of non-answerable questions. Besides photo quality, these questions and photos are often labeled as “unanswerable” due to lack of information in the image. In this case, external knowledge base such as [10] might help. Since in this project we are not building VQA models that gives strict answer, we consider this area out of scope of this study.

CHAPTER 3
METHOD

3.1 Initial Analysis on the VizWiz Dataset

We first conducted some basic analyses on the Vizwiz training dataset to see its distribution. There are 20,000 samples in total. Each sample contains 10 answers collected from crowd sources. Vizwiz categorized all samples into 4 types based on answers: number, yes/no, unanswerable, other. 'Number' type samples are mainly asking temperature, denomination or numbers in the image. 'Yes/no' type samples are mainly checking whether the user's conjecture is right or asking whether the light is on etc. Some samples are unanswerable due to vague images or images not containing enough information to answer the questions. Among 20,000 samples, 337 are 'number' type, 998 are 'yes/no', 6834 are 'unanswerable' and the rest 11,831 are 'other'.

For the main category 'other', we would like to see the top questions asked by users so as to learn their main concerns. Different users ask similar questions in their own words. Therefore, we first leverage word lemmatization and n-gram techniques to compress the questions. Our experiments show only n greater than 3 is suitable to learn the distribution of the questions and we tried $n = 4, 5, 6, 7$. Moreover, it is necessary to filter the polite expressions such as *could you please tell, thank you, please*. For the sake of accuracy, we manually inspect the compressed questions and list the top original questions in Table 3.1.

We can see most cases are object detection and then feature (color /flavor) recognition. Besides, Optical Character Recognition (OCR) tasks asking contents /name /title also occupies a high rate. The different tasks need different models. We choose to first focus on object detection only. In the following sections, we try to map the crowd-sourced answers in Vizwiz to systematic labels

Question	Frequency
What is it /this (item /product)?	3,931
What is in this box /can /bottle?	1,017
What color is this (item /shirt) / are these (pant) /my ...?	1,006
What kind of soup /coffee /soda /beer /chip /wine /dog /drink /tea is this?	401
What does this say?	250
What is the name of this product /book?	231
What flavor is this ...?	161
What is the expiration date?	117
For how long do I cook this in the microwave?	65
What is this a picture of?	59
What is the title of this book?	47
What is on the screen?	38

Table 3.1: Top questions in ‘other’ type samples

which can be used as ground truth to build machine learning /deep learning models.

3.2 Mapping VizWiz to ImageNet

Intuitively, to build a computer vision model on a downstream task such as this specific domain, we would use a pre-trained model like ResNet-152 that are trained on very large image dataset like ImageNet, and fine-tune on our images and annotations. So the first step towards building the model is to map VizWiz annotations, which consists of QA pairs, into simple ImageNet-1k labels.

Our attempts of mapping is illustrated in Figure 3.1. We first identify nouns through Part-of-Speech tagging, provided in the NLTK toolkit [7]. We then use WordNet [8] to find the corresponding synset id, which is also an unique identifier of the ImageNet categories.

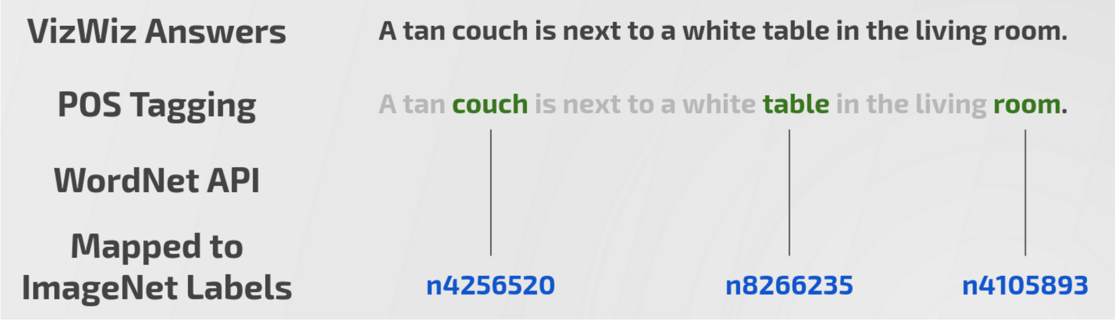
VizWiz Answers
A tan couch is next to a white table in the living room.
POS Tagging
A tan couch is next to a white table in the living room.
WordNet API
Mapped to ImageNet Labels
n4256520
n8266235
n4105893

Figure 3.1: Mapping VizWiz Answers into ImageNet Labels

Using this method, we are only able to map less than 100 instances from the VizWiz dataset. More on this would be clarified in the Discussion chapter.

3.3 Building New Labels for Object Detection

After the failure of mapping to ImageNet labels, we realized the necessity to build our own labels instead of using the messy categories defined in other tasks.

3.3.1 Single word version

We first filter out all non-object detection samples by checking the question field. Based on Table 3.1, we assume object detection questions start with *What is this, What is it*, *What's this*, *What's it*, *What's in*, *What is in*. Then we have 6,202 samples out of 20,000 training data.

For each selected sample, we use *TextBlob* to extract all noun phrases from 10 answers and then lemmatize each word and add it to our vocabulary. The vocabulary is sorted by each word frequency. There are 4,315 words in total and the threshold point - 1,000th word (i.e. wifi) appears 9 times in all answers. Table 3.2 shows the top 52 words and their corresponding frequency in answers.

Word	Frequency	Word	Frequency	Word	Frequency	Word	Frequency
chicken	910	sauce	689	bottle	620	bean	563
phone	547	mix	495	soup	484	beef	465
computer	435	coffee	380	cream	337	cup	334
green	333	cell	332	box	330	cheese	308
tomato	302	remote	282	white	282	control	280
water	273	pie	266	juice	261	butter	252
black	249	hand	246	food	242	frozen	224
pasta	214	chocolate	211	bag	208	red	207
apple	204	chip	197	pot	195	potato	192
spaghetti	191	hot	185	dinner	185	dog	184
tea	182	keyboard	181	peanut	181	dollar	178
screen	176	mushroom	176	cake	174	bill	173
paper	170	orange	168	noodle	168	lotion	166

Table 3.2: Top 52 words in vocabulary

To have an effectiveness evaluation of our vocabulary, we would like to see whether the image contains the object as claimed in answer. For each of the top 10 word, we display 20 images the answer of which contains the word. Figure 5.1 - Figure 5.8 in appendix part show the results.

We can see although the word is a key information in the images, it might not be the object. e.g. In Figure 5.1, there are a lot of chicken flavor food instead of a chicken. So are the cases for other food words. And in Figure 5.5, some images are actually accessories for a phone such as a phone cable. We realized we cannot cut a noun phrase into words. Instead, a noun phrase should be considered as a whole.

3.3.2 Noun phrase version

In the final refined version, we first use *TextBlob* to extract all noun phrases. And then use *nltk.pos_tag* to filter adjectives in the noun phrases and also lemmatize each token. Yet in the end, we concatenate the noun phrase and treat it as a whole and insert into our vocabulary. In this way, there are 8,106 noun phrases in our vocabulary. Table 3.3 shows top 50 noun phrases. From the difference between Table 3.2 and Table 3.2, we can infer that most of the top words in Table 3.2 are actually a part of a noun phrase and probably works as a modifier for the following noun e.g. 'chicken' in 'chicken flavor', 'cell' in 'cell phone'. Figure 5.9 - Figure 5.16 show the results for the top 10 noun phrases. We can clearly see Figure 5.9 is more accurate than Figure 5.5.

Word	Frequency	Word	Frequency
cell phone	281	control	267
bean	263	dollar bill	153
water bottle	152	tomato sauce	152
hand sanitizer	143	coca cola	117
computer mouse	116	coffee cup	94
kidney bean	84	orange juice	81
peanut butter	80	macaroni cheese	78
cream mushroom soup	72	computer keyboard	69
cake mix	67	coke	63
body lotion	62	bottle	61
coffee mug	60	dinner	60
cup	57	chicken noodle soup	57
cottage pie	57	pocket	55
spaghetti	51	pepper	50
phone	49	chicken dijon	49
chicken broth	48	chicken pot pie	47
rice	47	computer	46
potato chip	46	spaghetti meatball	46
computer screen	45	soy sauce	44
roast beef	41	i dont	40
ground beef	40	savoury beef rissole	40
beef broth	40	image	38
hamburger helper	38	pinto bean	37
pea	36	brownie mix	36
pot pie	35	chair	35
beef stroganoff	26	pepsi	26
air freshener	26	alarm clock	26
fire extinguisher	26	wine bottle	25

Table 3.3: Top 50 noun phrases in vocabulary

CHAPTER 4

DISCUSSION

4.1 ImageNet Mapping

There are a lot of potential reason that contributed to the failed attempt of mapping to ImageNet label. The main reason is that ImageNet is mainly an image classification dataset. Its categories, often natural species of plants and animals, are not designed to distinguish day-to-day objects that visually impaired people would need, like can and container labels.

4.2 Building New Label Set

Although by taking advantage of noun phrases instead of single noun, we refine our vocabulary, there is still a lot can be improved. For example, there can be various expression for one object. In Table 3.3, we see 'coca cola' and 'coke' are both among top noun phrases, yet they are the same thing. So are 'coffee mug' and 'coffee cup'. Synset by wordnet might be helpful to combine the noun phrases into one category.

On the other hand, we find the noun phrases can be of different hierarchy. e.g. 'coffee cup' is a subcategory under 'cup' which ranks a bit lower than the former in table. This issue is also in ImageNet, resulting in we building our own labels. And as we stated in previous section, utilizing synset by wordnet does not do a great favor.

4.3 Future Work

While we believe that our work on building a new label set of visual recognition for visually impaired people would inspire more future works on this domain, we believe that our work is limited to the number of training data and well-defined annotation. A large amount of photos taken by visually impaired people and manually labeled dataset by crowd-sourcing, would be necessary. For example the object detection dataset COCO has 245,496 annotated photo in total. Providing sufficient training data would allow more room for fine-tuning existing computer vision models, and more testing data would give an accurate benchmark for this task. Moreover, we find the image set provided by Vizwiz are of various sizes which add difficulty to build a deep learning model. Even though upsampling or downsampling can help solve the issue, it would inevitably add noises in the model and result in lower accuracy.

CHAPTER 5
CONCLUSION

In summary, this work provides insight to the VizWiz dataset and make contributions to mapping its crown-sourced answers to more systematic labels. Our label set serves as a foundation for more sophisticated datasets and neural network models on this domain.

BIBLIOGRAPHY

[1] Jan Balata, Zdenek Mikovec, and Lukas Neoproud. Blindcamera: Central and golden-ratio composition for blind photographers. In *Proceedings of the Mulitimedia, Interaction, Design and Innnovation*, MIDI '15, pages 8:1–8:8, New York, NY, USA, 2015. ACM.

[2] Jeffrey Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samuel White, and Tom Yeh. Vizwiz: Nearly real-time answers to visual questions. pages 333–342, 01 2010.

[3] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[5] Chandrika Jayant, Hanjie Ji, Samuel White, and Jeffrey P. Bigham. Supporting blind photography. In *The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '11, pages 203–210, New York, NY, USA, 2011. ACM.

[6] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[7] Edward Loper and Steven Bird. NLTK: the natural language toolkit. *CoRR*, cs.CL/0205028, 2002.

[8] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November 1995.

[9] Marynel Vázquez and Aaron Steinfeld. An assisted photography framework to help visually impaired users properly aim a camera. *ACM Trans. Comput.-Hum. Interact.*, 21(5):25:1–25:29, November 2014.

[10] Qi Wu, Chunhua Shen, Anton van den Hengel, Peng Wang, and Anthony R. Dick. Image captioning and visual question answering based on attributes and their related external knowledge. *CoRR*, abs/1603.02814, 2016.

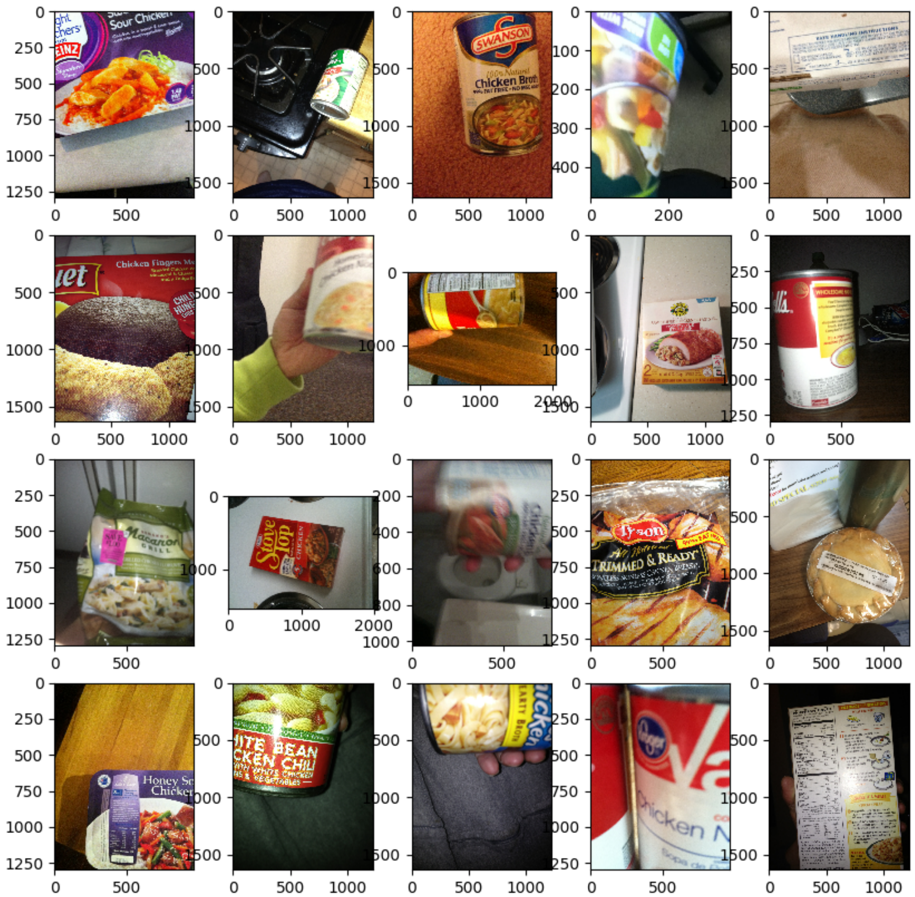

Figure 5.1: 910 Images containing ‘Chicken’ in answers

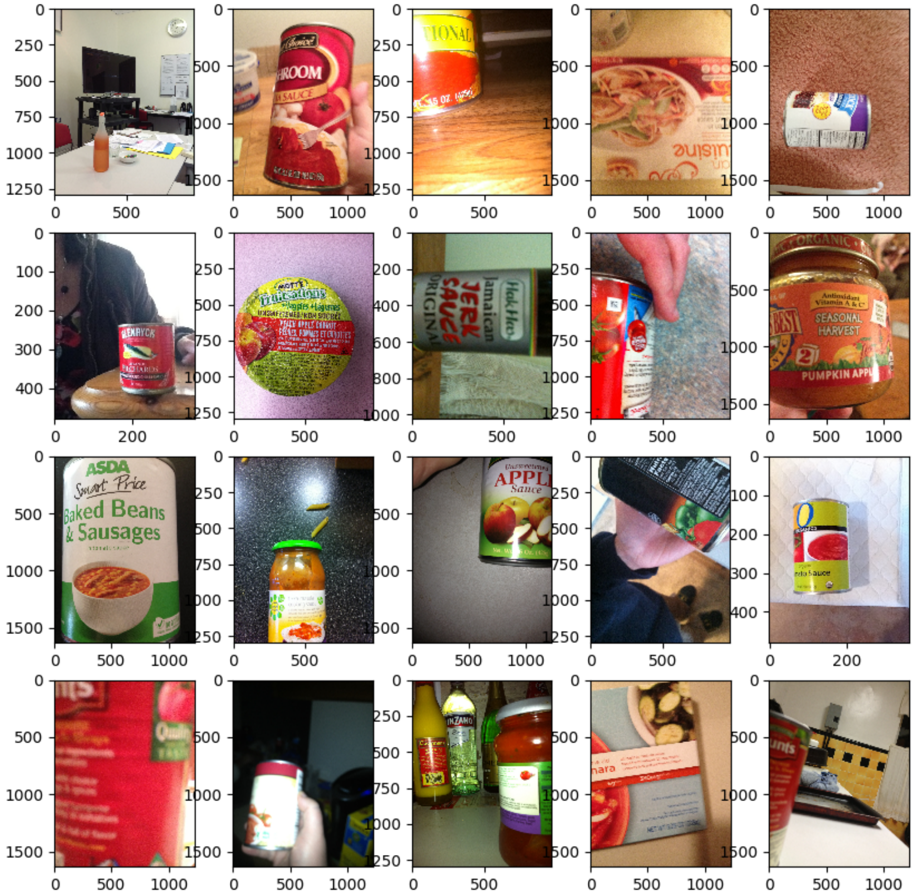

Figure 5.2: 689 Images containing ‘Sauce’ in answers

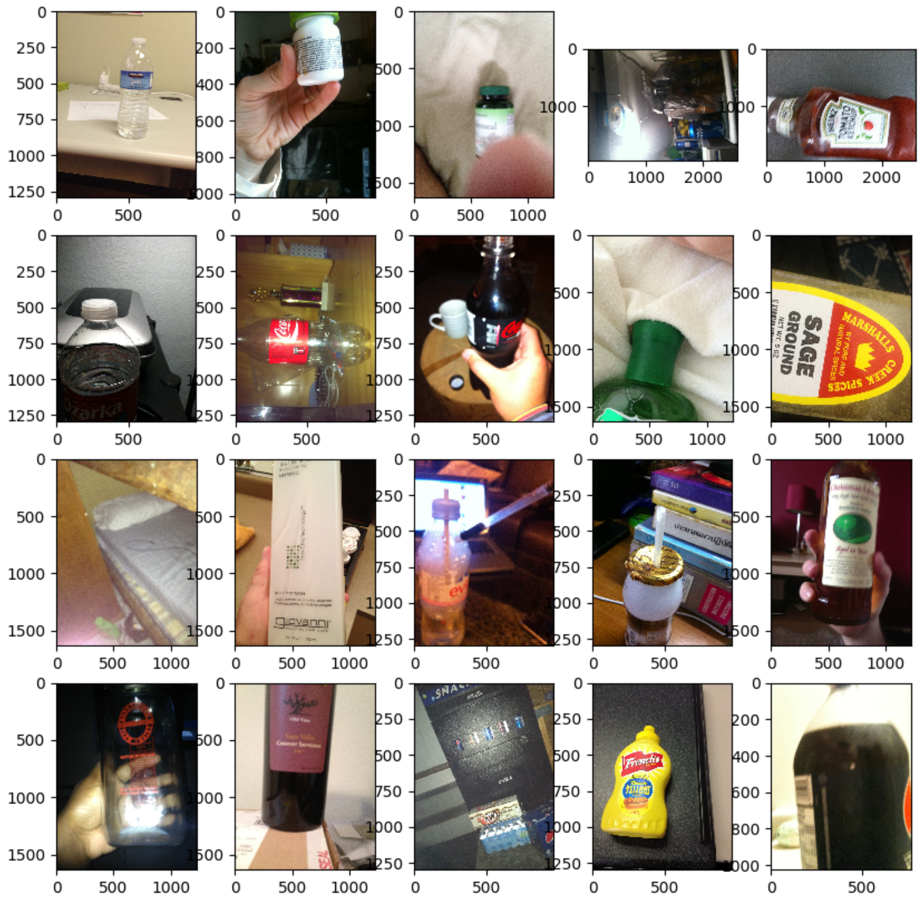

Figure 5.3: 620 Images containing 'Bottle' in answers

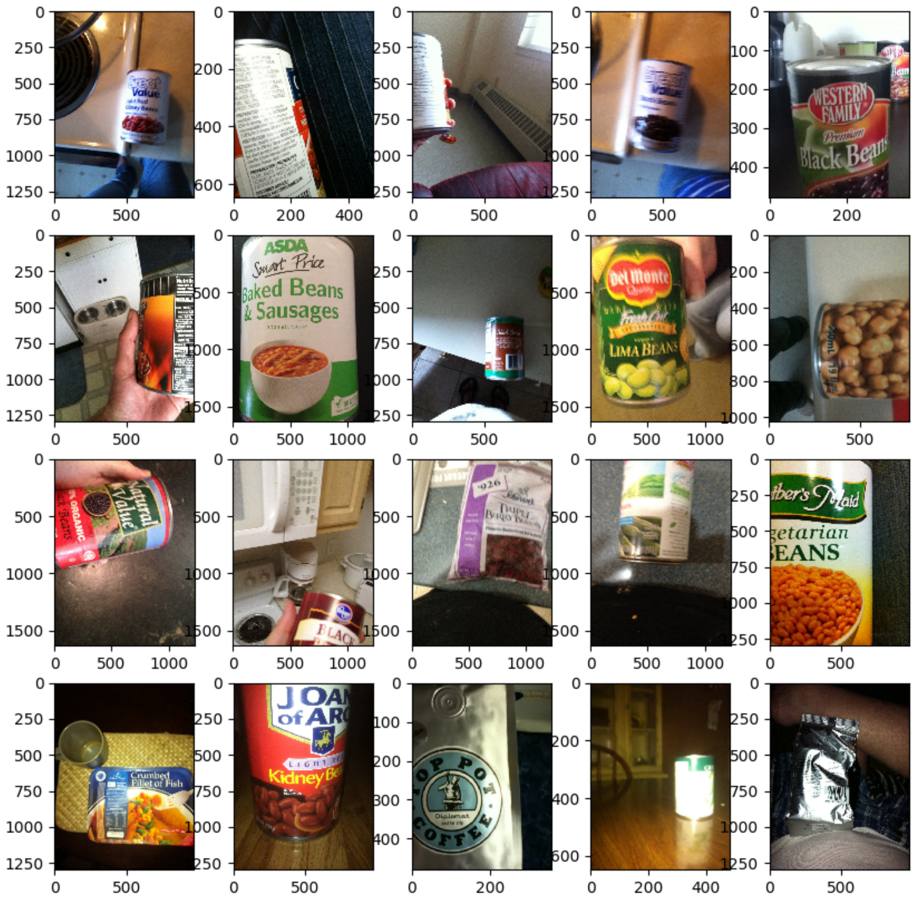

Figure 5.4: 563 Images containing 'Bean' in answers

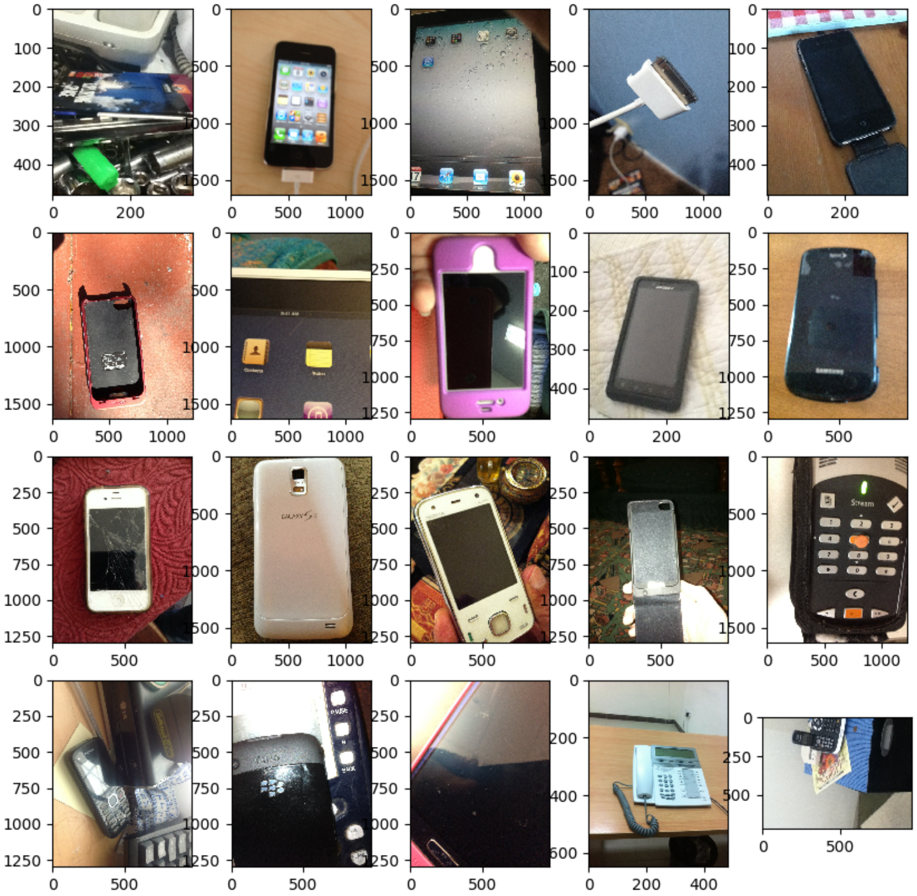

Figure 5.5: 547 Images containing 'Phone' in answers

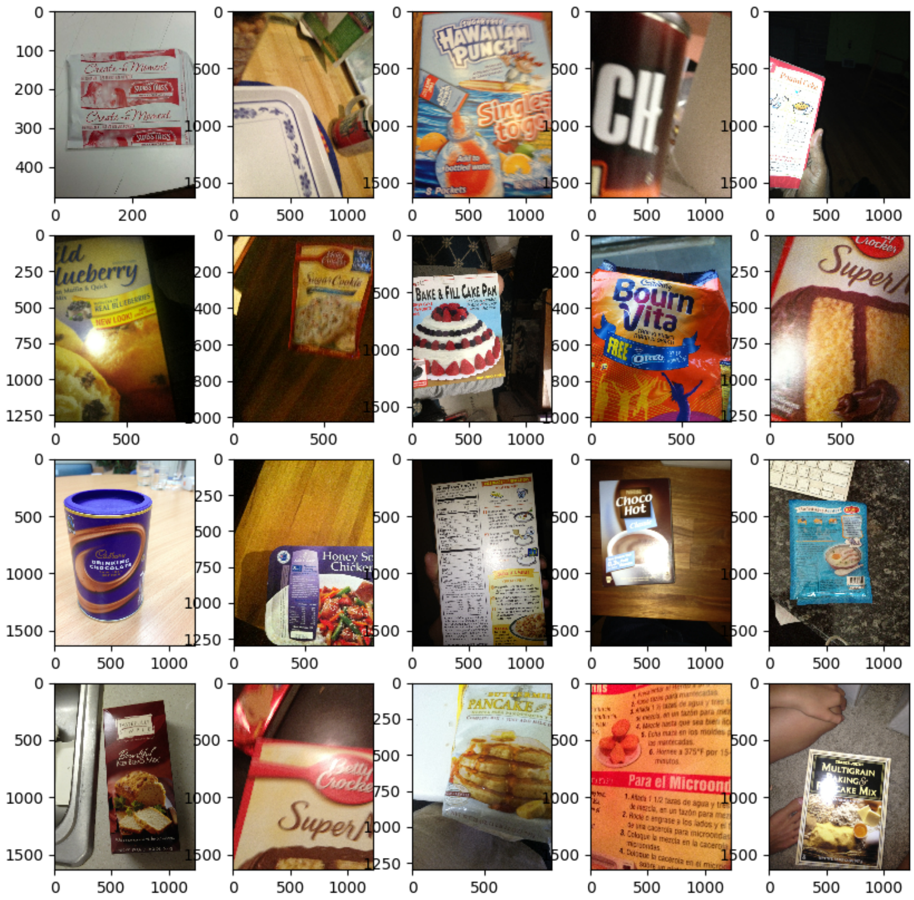

Figure 5.6: 495 Images containing ‘Mix’ in answers

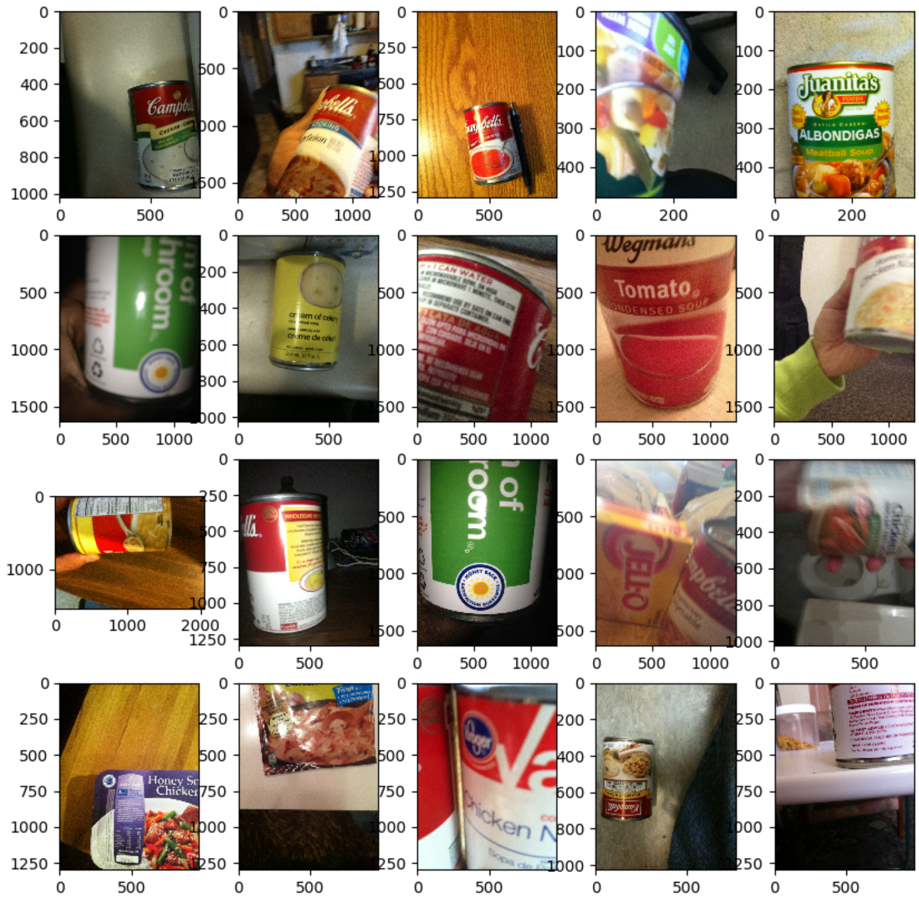

Figure 5.7: 484 Images containing 'Soup' in answers

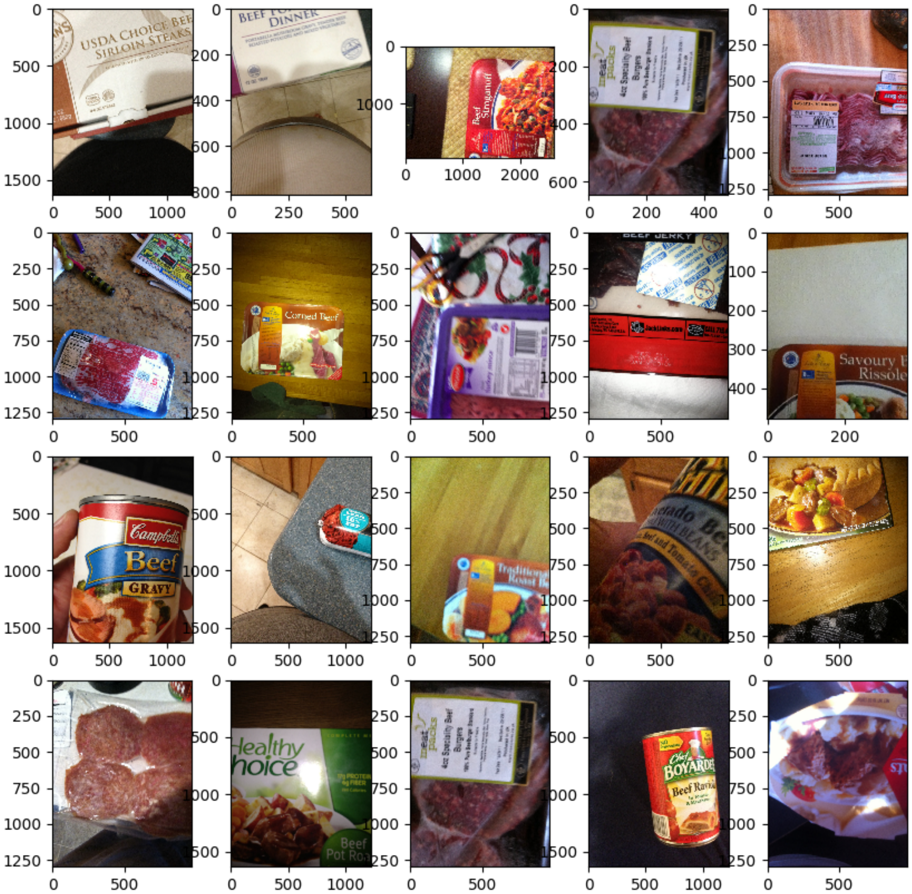

Figure 5.8: 465 Images containing ‘Beef’ in answers

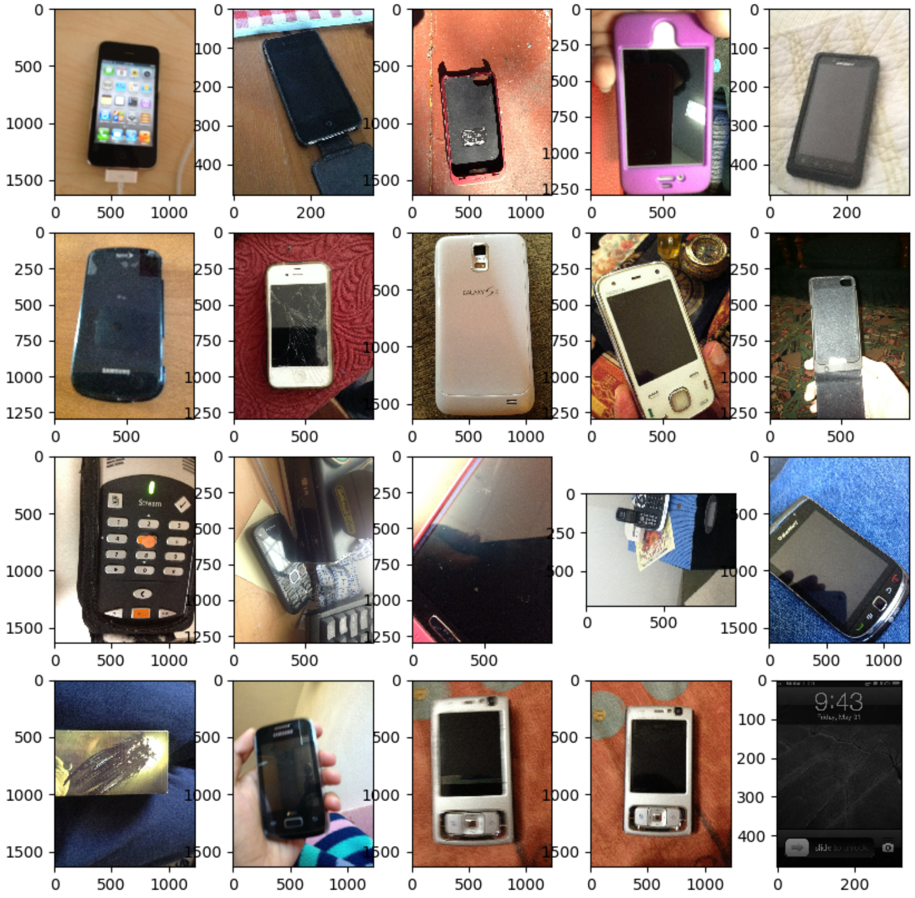

Figure 5.9: 114 Images containing 'Cell Phone' in answers

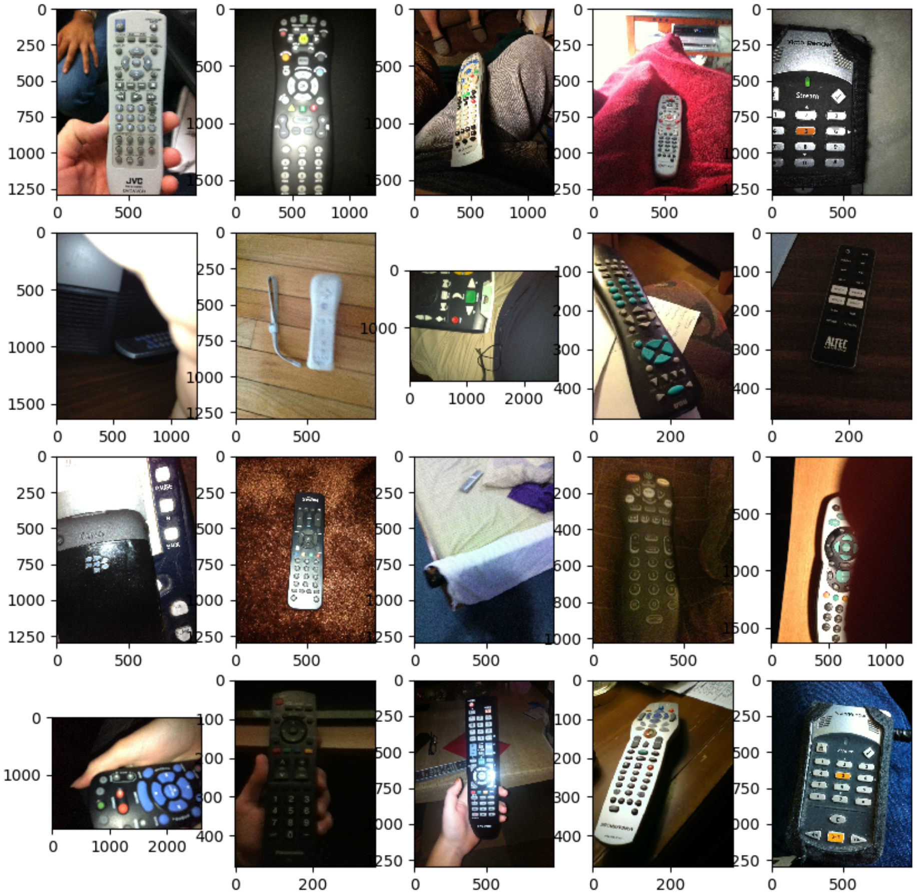

Figure 5.10: 83 Images containing 'Control' in answers

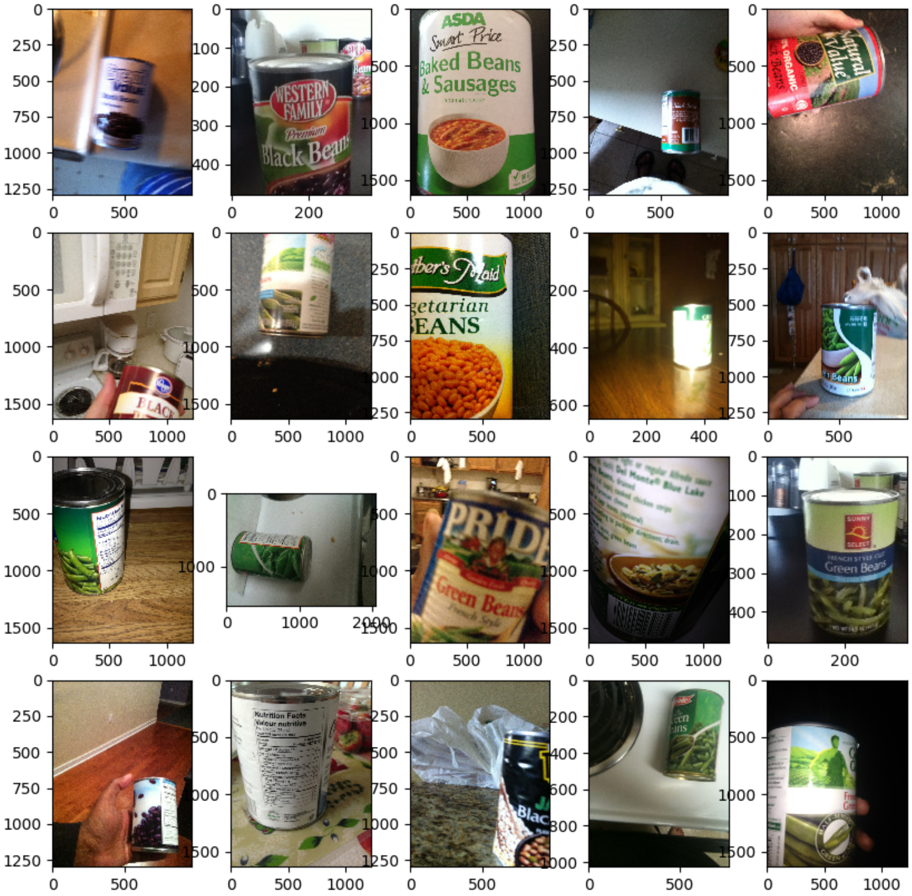

Figure 5.11: 49 Images containing 'Bean' in answers

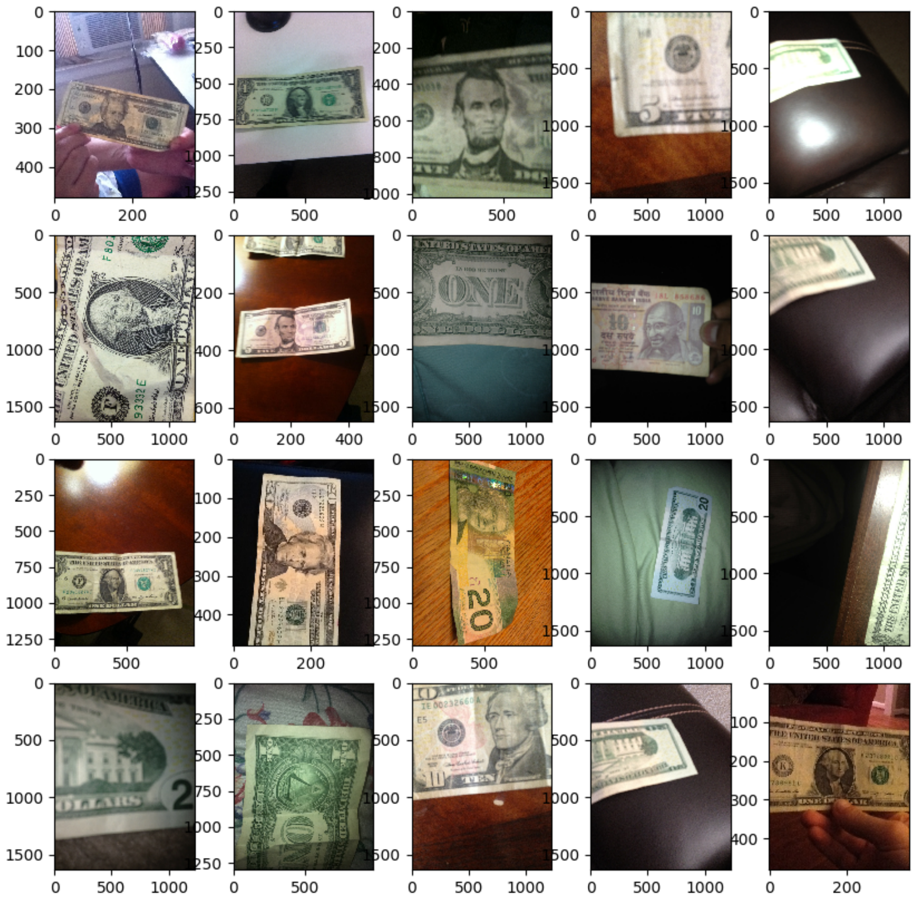

Figure 5.12: 37 Images containing 'Dollar Bill' in answers

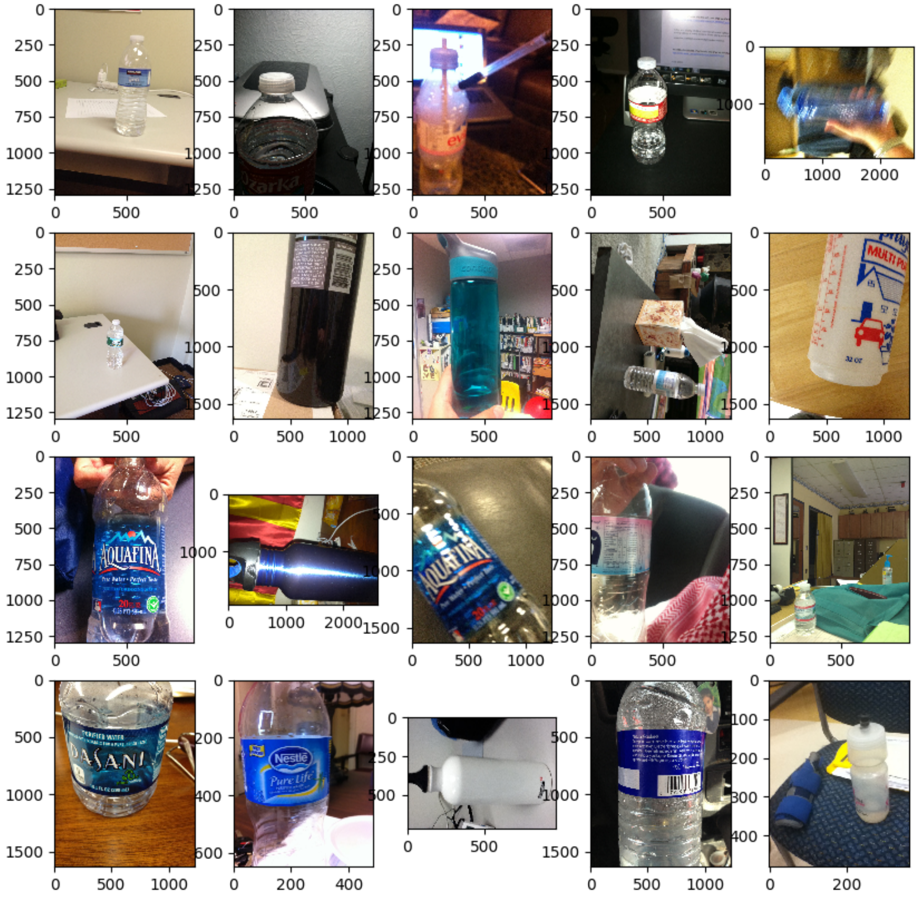

Figure 5.13: 43 Images containing 'Water Bottle' in answers

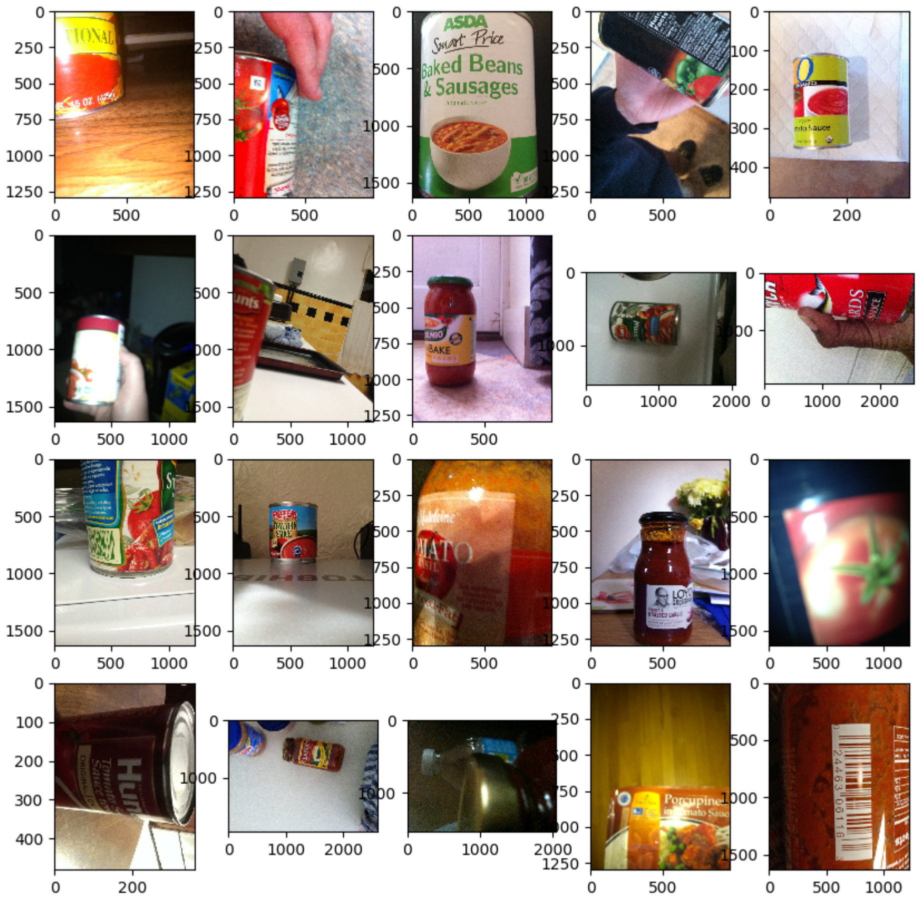

Figure 5.14: 48 Images containing 'Tomato Sauce' in answers

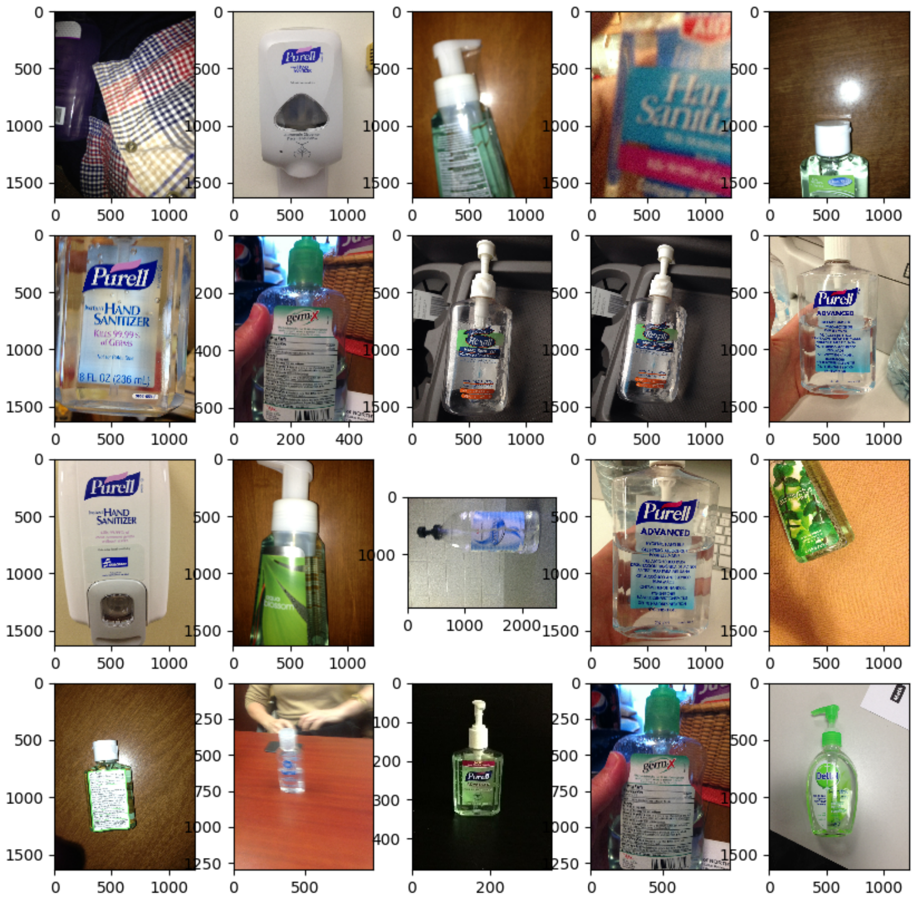

Figure 5.15: 27 Images containing 'Hand Sanitizer' in answers

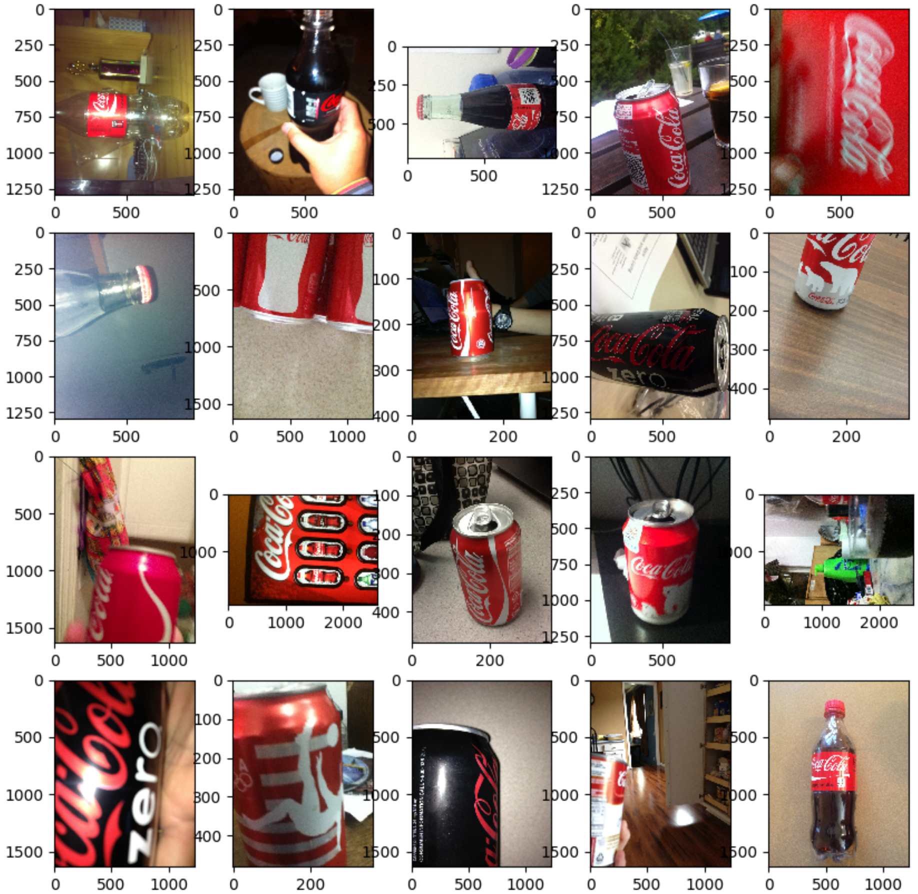

Figure 5.16: 27 Images containing 'Coca Cola' in answers