CS395T Grounded NLP Project: Data Augmentation in Unanswerable VQA

Ryo Kamoi
ryokamoi@utexas.edu

Jie Hao Liao
jhliao@utexas.edu

Abstract

We study data augmentation for answerability classification for visual question answering (VQA). Dataset bias is a major problem in many VQA dataset, and models can learn to answer using only questions or only images, causing problems in real world applications. Answerability classification to detect unanswerable question is a way to alleviate this problem and improve the safety and robustness of VQA systems. Answerability classification is often trained by supervised learning on unanswerable questions in training data, but many datasets do not contain sufficiently large and diverse unanswerable data and can be affected by dataset biases. We create unanswerable data by assigning a randomly chosen question to an image. Our experiments show that our strategy does not provide any significant improvement, but we show some results motivating further work in data augmentation for answerability classification in VQA.

1 Introduction

Detection of unanswerable questions is necessary for improving the safety of question answering systems. In this work, we study data augmentation for answerability classification for visual question answering (VQA). It has been observed that VQA models have often suffered from language biases (Manjunatha et al., 2019), and they sometimes can answer questions without using image inputs. For example, their answer to the question “What is the color of the grass?” is usually “Green”, even when there is no grass in the provided image. To avoid this problem, we can use a model to judge if a question about an image can be answered.

However, there are very few datasets (Davis, 2020) that contains unanswerable VQA questions for training. In this project, we propose to synthesize unanswerable image-question pairs as data augmentation for training to improve the safety of VQA models.

We mainly use VizWiz dataset (Gurari et al., 2018), which contains realistic and possibly unanswerable questions from blind people. First, to show that data augmentation is required, we show results on the model trained on a subset of data. Next, we apply a simple random swap data augmentation, which assigns a question to an image randomly. Our experiments show that this strategy is not effective. As future work, we propose to make a more difficult negative pair by assigning a question or an image similar to the original one.

2 Background

2.1 Problem Formulation

In our answerability task, we aim to predict the probability $P(y|x_{img}, x_{txt})$ given an image x_{img} and a question x_{txt} where y is the event that the image-question pair is answerable. That is, the image can be used to answer the question. We deem an image-question pair as unanswerable if the image cannot be used to answer the question. Since a image-question pair can either be answerable or unanswerable, the answerability task is a binary classification task.

2.2 UNITER

UNITER (Chen et al., 2019) is a large-scale pretrained model for joint visual and textual understanding based on Transformers (Vaswani et al., 2017). First, UNITER encodes image regions (visual features and bounding box features) and text into a common embedding space with image and text embedders. Then, a Transformer module is applied to learn contextualized embeddings. The model is pretrained on self-supervised learning tasks: (i) masked language modeling conditioned on image, (ii) masked region modeling conditioned on text, (iii) image text matching, and (iv) word region alignment.

We applied a feed-forward network on the output of the [CLS] token from UNITER for binary classification on the answerability task. Note that this [CLS] output is pre-trained using the Image-Text Matching objective as proposed in Section 3.2 of (Chen et al., 2019). We fine-tuned the feed-forward network on the binary cross-entropy objective:

$$\begin{aligned}\mathcal{L}_{std} = -E_{(y,x_{img},x_{text})\sim\mathcal{D}} \\ [\alpha \cdot (y \log s_\theta(x_{img}, x_{text}) + \\ (1-y)\log s_\theta(x_{img}, x_{text}))]\end{aligned} \quad (1)$$

where x_{img}, x_{text} denote an image-question pair, y denotes if the image-question pair is answerable, and $s_\theta(\cdot,\cdot)$ denotes the score output of feed-forward network. Since there is a data imbalance of possibly more answerable questions than unanswerable questions, α is a hyperparameter that can be used to weigh the importance of unanswerable questions in the objective.

2.3 VizWiz

VizWiz (Gurari et al., 2018) is a VQA dataset where images and questions are sourced from blind people. The dataset contains blurry and poor quality images as well as unanswerable questions since the people who took the images cannot actually see the images. For each question, answers are crowd-sourced from 10 Amazon Mechanical Turks. If at least half of the crowdsourced answers for a visual question is ”unanswerable” or ”unsuitable image”, the question is considered unanswerable. 28.63% of questions in the dataset are unanswerable.

Since the answerability of a question is part of the 10 answers for a question, we propose to mark a question that contains ”unanswerable” or ”unsuitable image” in at least half of the answers as an unanswerable question. This will group each visual question into two classes, answerable and unanswerable. We propose to use this to train models on a binary classification task for answerability.

2.4 Synthetic VQA

VQA 1.0 For a simple baseline in data augmentation, we exchanged images for each question in the VQA 1.0 dataset (Ray et al., 2016). This will generate image-question pairs that are possibly fully irrelevant, but is nonetheless a good starting point. We checked if an answerability model trained on such simple augmentation can improve its performance.

QRPE The Question Relevance Prediction and Explanation (QRPE) dataset (Mahendru et al., 2017) is a VQA dataset that contains a tuple of a relevant image and a slightly irrelevant image for a question. The authors use the premises of a visual question to find images that have relevant premises but cannot be used to answer the visual question. The dataset is created using MS COCO images (Lin et al., 2014), which is the same source for VQA 1.0.

2.5 VILLA

VILLA (Gan et al., 2020) applies adversarial training to improve vision-and-language representation learning. To adopt the “free” adversarial training which enables large-scale training, they propose to perform adversarial training in the embedding space, instead of adding adversarial perturbations on image pixels and textual tokens. Instead of random perturbations, the adversarial perturbations are updated through gradients from the training objective. VILLA achieved state-of-the-art across six visual-and-language tasks. Specifically related to this work, VILLA improves the single-model performance of UNITER-large from 74.02 to 74.87 on VQA.

3 Methodology

Original Dataset In the original VizWiz dataset, 28.63% of the questions are unanswerable. To show the necessity of data augmentation, we evaluated model performance after fine-tuning it on different amounts of the original training data. For example, we compared the results of training with 0%, 1%, 5%, 10%, 25%, 50%, 75%, and 100% of the unanswerable questions in the original training data. We also fine-tuned our model on only the text modality and only the image modality to see if the answerability of the questions in the dataset are biased towards one modality.

Data Augmentation (Mix) We mixed the original and 0%, 1%, 5%, 10%, 25%, 50%, 75%, and 100% of the synthetic datasets to improve the performance of answerability classification.

VILLA We used the adversarial training as proposed in VILLA (Gan et al., 2020) to generate artificial data augmentation during training. VILLA proposed a method to update the perturbations through gradients during training. We also evaluated the performance on using perturbations initialized from random uniform distributions (without gradient updates) to fine-tune our model instead of gradient-updated perturbations. We applied the perturbations on the text modality alone, on both the text and image modalities, and on alternating between the text and image modalities.

3.1 Quantitative Evaluation

We used the accuracy of answerability classification for evaluation. We define the answerable questions accuracy as accuracy on only the answerable questions and unanswerable questions accuracy as accuracy on only the unanswerable questions. We report our model's F1 score on the answerable questions, the answerable questions accuracy, and the unanswerable questions accuracy.

We define balanced accuracy as the mean of the answerable questions accuracy and the unanswerable questions accuracy. Since a model can just answer everything as answerable on the VizWiz dataset and achieve about 70% in accuracy, the balanced accuracy is a good metric for measuring accuracy without bias towards classifying answerability or unanswerability too often. We report this to show that our model is not biased towards predicting only one particular class.

3.2 Qualitative Evaluation

For correctness in the synthetic datasets, we propose to sample a subset of the visual questions and check if the question is truly unanswerable given the image. We will also sample a subset of wrong answers by our model to find the limitations of our model.

4 Implementation

4.1 Datasets

VizWiz Since there are already answerability labels in the VizWiz dataset which marked if a question is unanswerable or contains an unsuitable image, we directly used them as targets in training our models. We took 2000 examples from the VizWiz train set as validation for tuning hyperparameters. The actual VizWiz validation set is used as the test set for measuring performance on all models.

VQA 1.0 For the VQA dataset, we randomly picked two questions in the dataset with different image IDs and swapped their image IDs so that the questions is probably unanswerable. We continued this procedure until nearly half of the dataset contains unanswerable questions. We labeled the questions with swapped image IDs as unanswerable and the questions without swapped image IDs as answerable.

QRPE For the QRPE dataset, each question contains a relevant image that can be used to answer the question and an irrelevant image that cannot be used to answer the question. We labeled the question with the relevant image as answerable and question with the irrelevant image as unanswerable. Thus, there are two examples per question from QRPE.

The statistics of answerable and unanswerable questions for all datasets can be found in Table 3 in the Appendix.

4.2 UNITER

We expanded on the official implementation of UNITER as the model for solving answerability.[1] Essentially, we modified the UNITER VQA model and training code to do a binary classification of whether or not a question is answerable or not instead of the true question answering task. We set the threshold of a visual question to be answerable if the classification output is 0.5 in probability or greater, and we used $\alpha = 3.5$ in equation 1 for the fine-tuning objective. Our implementation can be found on GitHub.[2]

We used both the base UNITER model and the large UNITER model in our experiments. Our model and training hyperparameters can be found in Table 1 in the Appendix. In general, we used the default hyperparameter configurations for fine-tuning on a VQA task in the official UNITER repository as the hyperparameter configurations for our answerability task's experiments.

4.3 VILLA

For the VILLA training procedure, we used code from the official VILLA repository since the authors of VILLA built VILLA on top of UNITER.[3] Our objective with VILLA for the answerability task follows from the objective in Section 3.3 of

[1] https://github.com/ChenRocks/UNITER

[2] https://github.com/liaojh1998/unans-vqa/tree/qrpe_prepro/Jay

[3] https://github.com/zhegan27/VILLA/

(Gan et al., 2020):

$$\min_{\theta} E_{(x_{img}, x_{txt}, y) \sim \mathcal{D}} [\mathcal{L}_{std}(\theta) + \mathcal{R}_{at}(\theta) + \beta \cdot \mathcal{R}_{kl}(\theta)] \quad (2)$$

where $\mathcal{L}_{std}(\theta)$ is the standard binary cross entropy loss, $\mathcal{R}_{at}(\theta)$ is the binary cross entropy loss using adversarial examples, and $\mathcal{R}_{kl}(\theta)$ is the Kullback-Leibler Divergence loss for matching the confidence level of the answerability prediction with the true answerability distribution. We set $\beta = 1.5$ as done in the default hyperparameters for fine-tuning a VQA task using VILLA. A list of our other hyperparameters can be found in Table 2 in the Appendix.

4.4 Resources

We used the UT HTCondor cluster[4] for computational resources. There are several `eldar` machines on the cluster that contains GPUs with 12GB of graphics memory. We also used computational resources from TACC[5]. The `maverick2` hosts on TACC contains GPUs with 16GB and 32GB of graphics memory. Note that 12GB of graphics memory sufficed for fine-tuning the base UNITER model, so most of our experiments are conducted on the UT HTCondor cluster.

5 Results

5.1 VizWiz

We found the default hyperparameters for fine-tuning UNITER on VQA worked quite well for fine-tuning on the answerability task. The base UNITER model generally converged in accuracy after 6000 fine-tuning steps. The base UNITER model seems to achieve similar performance as with the large UNITER model, so we used the base UNITER model for all of our other experiments. These results are reflected in Tables 4 and 7 in the Appendix.

In general, fine-tuning on information from one modality is unable to perform as well as fine-tuning on information from both image and text modalities. UNITER fine-tuned on only the text modality achieved 67.77% in accuracy, on only the image modality achieved 73.74% in accuracy, and on both modalities achieved 81.45% in accuracy. Information on other accuracy can be found in Table 5 in the Appendix. This demonstrates that a certain amount of image-question pairs in the VizWiz dataset cannot be determined as answerable or unanswerable without checking both modalities.

Fine-tuning on a subset of the VizWiz train set generally achieved less in accuracy than fine-tuning on all of VizWiz train set. UNITER achieved only 81.45% in accuracy after fine-tuning on 100% of VizWiz's train set, and a summary of accuracy information can be found in Table 6 in the Appendix. Thus, data augmentation can be helpful for improving performance.

Lastly, UNITER usually achieves better performance in answerable questions accuracy than unanswerable questions accuracy, but it does not bias towards predicting only answerable or unanswerable for classification.

5.2 Synthetic VQA

VQA 1.0 (Random Swap) We augmented the full VizWiz train set with multiple different subsets of the VQA 1.0 train set, and fine-tuned UNITER on the augmented sets. Generally, using more augmentation decreased validation accuracy on VizWiz validation set, especially the answerable questions accuracy, but increased the validation accuracy on the VQA 1.0 validation set. This is reflected in Tables 8 and 9 in the Appendix. Thus, there may be harmful examples in the VQA 1.0 train set that are not beneficial for improving answerability accuracy on the VizWiz dataset.

It is also important to note that VQA 1.0 train set contains 10 times more data than VizWiz train set. Thus, our UNITER model may be biased in fine-tuning to improving answerability more on VQA 1.0. However, using a similar amount of VQA examples and VizWiz examples is still detrimental to performance. Thus, we propose to not use this kind of synthetic data augmentation in future work.

QRPE We augmented the full VizWiz train set with multiple different subsets of the QRPE train set, and fine-tuned UNITER on the augmented sets. Generally, the performance on the VizWiz validation set did not change much on varying the amount of data augmentation. According to Table 12 in the Appendix, sometimes there are improvements, and sometimes there are not. We propose the changes in performance may be due to better random initialization, or that in the QRPE dataset, certain examples may be helpful and certain examples may

[4] https://www.cs.utexas.edu/facilities/documentation/condor

[5] https://www.tacc.utexas.edu/systems/maverick2

be harmful to VizWiz.

Again, found similarly with other datasets, using more augmentation of the QRPE train set improves performance on the QRPE validation set. This is reflected in Table 13 in the Appendix.

5.3 VILLA

Perturbations on the Text Modality In general, perturbations on only the text modality seemed to increase answerable questions accuracy, but decrease unanswerable questions accuracy. This is the case for all perturbations initialized randomly from uniform distributions of values $[0.01, 0.001, 0.0001, 0.00001]$ without any gradient updates.

Perturbations on Both Modalities It seems that perturbations on both modalities seemed to improve performance on the VizWiz validation set by a tiny bit, like about 0.2% or so in accuracy. This is reflected in Table 16, where using random perturbations achieved an accuracy of 81.66% and using gradient-updated perturbations achieved an accuracy of 81.59%, as opposed to an accuracy of 81.45% without using any perturbations. Also, alternating perturbations in each modality may do better than adding perturbations to both modalities at once. Since we only have one sample of this, this claim needs to be verified further.

5.4 Qualitative Analysis

After checking multiple qualitative examples in our base UNITER model, we are unable to come to any conclusions on why a certain image would be considered answerable or unanswerable by the model, and when would a model be correct or incorrect. We found that our model predict confidently and correctly on answerability for questions that usually have a unanimous vote for a clear answer, and incorrectly for questions that have different votes for different answers. Some examples can be found in the Appendix in Figures 2 and 1.

6 Related Works

6.1 Dataset Biases in VQA

Dataset biases have been observed in multiple VQA datasets. It has been reported that models only using questions or image inputs are are not significantly inferior to models using both questions and images on VQA dataset (Agrawal et al., 2016). Jabri et al. (2016) shows that state-of-the-art VQA systems is not significantly better than that of systems designed to exploit dataset biases.

Multiple approaches have been proposed to alleviate the problems of dataset biases in VQA. Zhang et al. (2016) proposes a way to balance the labels by creating synthetic dataset. However, it is difficult to apply this strategy to real images. Goyal et al. (2017) carefully design data collection strategy to make a balanced dataset of real images, but it is difficult to apply it to more complex and large scaled dataset. RUBi (Cadene et al., 2019) reduces the importance of the most biased examples to remove the influence by the dataset biases.

Another approach towards the dataset biases is to provide interpretable explanation about models (Manjunatha et al., 2019).

6.2 Hard Negative Mining

Our work is related to a technique called hard negative mining, selecting those examples for which the detector triggers a false alarm (Shrivastava et al., 2016). Hard negative mining is motivated by the observation in many datasets that they contain an overwhelming number of easy examples and a small number of hard examples.

Different from hard negative mining in prior work, in this work, we create hard negative sample via data augmentation to alleviate the effect of dataset biases.

6.3 Privacy Datasets

VizWiz-Priv (Gurari et al., 2019) is a VQA dataset that contains images taken by blind people that may contain private information. Regions in the images that contain private information are masked out for privacy concerns. While a region may be masked out, baseline models demonstrate that it is still possible to realize these regions are private given the context around it. The dataset also contain randomly masked regions in other images to avoid a model to bias on indicating a question is unanswerable because its image is masked. In our stretch goals, we propose to use this dataset for data augmentation and for testing answerability models for privacy concerns.

VISPR (Orekondy et al., 2017) is another dataset that contains attributes on images that can present a certain privacy risk. We propose to form synthetic questions on these attributes for a VQA task as data augmentation to improve model performances on the VizWiz-Priv and VizWiz dataset.

For both datasets, we formulate a visual question as answerable if it does not contain private information, and unanswerable if it contains private information.

7 Conclusion

We propose to apply data augmentation for answerability classification in visual question answering (VQA). We create unanswerable data by assigning a randomly chosen question to an image. Our experiments on VizWiz dataset show that our method does not provide any significant improvement. However, we show that increasing unanswerable questions data will improve the detection performance, motivating further work in data augmentation.

In future work, we expect that a better data augmentation method will improve the performance for answerability classification. Our current method assigns a question to an image randomly, resulting in an obvious mismatch that will not improve models so much. We can try creating more difficult image-question pairs by picking up a similar image or question to the original one.

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C Lawrence Zitnick, Dhruv Batra, and Devi Parikh. 2016. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Remi Cadene, Corentin Dancette, Hedi Ben-Younes, Matthieu Cord, and Devi Parikh. 2019. RUBi: Reducing Unimodal Biases for Visual Question Answering. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. UNITER: UNiversal Image-TExt Representation Learning. *ECCV 2020*.

Ernest Davis. 2020. Unanswerable Questions About Images and Texts. *Frontiers in Artificial Intelligence*, 3(July):1–10.

Zhe Gan, Yen Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. *arXiv*, (NeurIPS):1–13.

Yash Goyal, Tejas Khot, Aishwarya Agrawal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Danna Gurari, Qing Li, Chi Lin, Yinan Zhao, Anhong Guo, Abigale Stangl, and Jeffrey P. Bigham. 2019. Vizwiz-PRIV: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:939–948.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. VizWiz Grand Challenge: Answering Visual Questions from Blind People. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3608–3617.

Allan Jabri, Armand Joulin, and Laurens Van Der Maaten. 2016. Revisiting Visual Question Answering Baselines. *arXiv preprint arXiv:1606.08390*.

Tsung-yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C Lawrence Zitnick, and Piotr Dollár. 2014. Microsoft COCO: Common Objects in Context. *ECCV 2014*.

Aroma Mahendru, Viraj Prabhu, Akrit Mohapatra, Dhruv Batra, and Stefan Lee. 2017. The promise of premise: Harnessing question premises in visual question answering. *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 926–935.

Varun Manjunatha, Nirat Saini, and Larry S Davis. 2019. Explicit Bias Discovery in Visual Question Answering Models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. 2017. Towards a Visual Privacy Advisor: Understanding and Predicting Privacy Risks in Images. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-Octob(i):3706–3715.

Arijit Ray, Gordon Christie, Mohit Bansal, Dhruv Batra, and Devi Parikh. 2016. Question relevance in VQA: Identifying non-visual and false-premise questions. *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 919–924.

Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. 2016. Training Region-based Object Detectors with Online Hard Example Mining. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Conference on Neural Information Processing Systems (NIPS)*.

Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, Devi Parikh, and Virginia Tech. 2016. Yin and Yang: Balancing and Answering Binary Visual Questions. *arXiv preprint arXiv:1511.05099*.

A Appendices

All models, datasets, and results information can be found here.

B Hyperparameters

Tables 1, 2, 3 contains information for each model, training, and dataset used.

Parameter	Base UNITER	Large UNITER
Max Text Tokens	60	60
RPN BB Keep Threshold	0.2	0.2
Max RPN BBs	100	100
Min RPN BBs	10	10
Training Batch Size	5120	3072
Optimizer	AdamW	AdamW
Learning Rate	8e-5	5e-5
Optimizer Betas	0.9, 0.98	0.9, 0.98
Learning Rate Warm Up Steps	600	500
Gradient Accumulation Steps	5	4
Fine-tune Steps	6000	5000
Dropout	0.1	0.1
Weight Decay	0.01	0.01
Gradient Norm	2.0	2.0
Unanswerable Questions Weight	3.5	3.5
Answerable Question Threshold	0.5	0.5
Use 16-bit Floating Point Precision	True	True

Table 1: UNITER Hyperparameters. RPN stands for Region Proposal Network and BB stands for Bounding Boxes.

Parameter	Random Perturbations	Gradient Updated Perturbations	Gradient Updated Perturbations On Alternating Modalities
Initial Perturbation Values Distribution	Uniform	Zero	Zero
Training Batch Size	5120	2560	2560
Gradient Accumulation Steps	5	3	3
Perturbations Update Steps	N/A	3	6
Text Modality Perturbations Learning Rate	N/A	1e-3	1e-3
Image Modality Perturbations Learning Rate	N/A	1e-3	1e-3
KL Objective Weight	1.5	1.5	1.5

Table 2: VILLA Training Hyperparameters.

Dataset	Total Examples	Answerable Examples	Unanswerable Examples
VizWiz Train	18509	13427	4996
VizWiz Train Validation	2000	1464	536
VizWiz Validation	4319	2934	1385
VQA Train (Random Swap)	248349	150749	97600
VQA Validation (Random Swap)	121512	73766	47746
QRPE Train	70972	35486	35486
QRPE Validation	36850	18425	18425

Table 3: Number of examples in each dataset. The VizWiz train set is split into a train set for training and a validation set for validation during training. The actual VizWiz validation set is not used during training to avoid overfitting on it. We report accuracies on the VizWiz validation set, not the train validation set.

C Training on VizWiz Only

Tables 4, 5, 6, 7 have results about the Base UNITER model fine-tuned only on the VizWiz train set. All models in table 5 and 6 are fine-tuned for 6000 steps.

D Training on VizWiz with VQA (Random Swap)

Tables 8, 9 have results about the Base UNITER model fine-tuned on the VizWiz train set combined with some subset of the VQA train set. All models presented are fine-tuned for 6000 steps.

Fine-tuning Steps	Accuracy	Answerable Questions Accuracy	Unanswerable Questions Accuracy	Balanced Accuracy	Answerable Questions F1 Score
500	78.68	81.97	71.70	76.83	0.8393
1000	78.68	81.97	71.70	76.83	0.8393
1500	79.81	85.92	66.86	76.39	0.8526
2000	80.62	85.62	70.04	77.83	0.8572
2500	80.67	86.43	68.45	77.44	0.8586
3000	80.67	89.37	62.24	75.80	0.8626
3500	81.04	88.75	64.69	76.72	0.8641
4000	80.48	85.24	70.40	77.82	0.8558
4500	81.11	89.60	63.10	76.35	0.8657
5000	81.57	89.40	64.98	77.19	0.8682
5500	81.18	88.48	65.70	77.09	0.8646
6000	81.45	88.07	67.44	77.75	0.8658

Table 4: Accuracies on the VizWiz validation set at different fine-tuning steps. Generally, accuracy improved upon training the model further and plateaued around 5000 epochs or so.

Modality	Accuracy	Answerable Questions Accuracy	Unanswerable Questions Accuracy	Balanced Accuracy	Answerable Questions F1 Score
Text Only	67.77	75.80	50.76	63.28	0.7616
Image Only	73.74	87.46	44.69	66.08	0.8190
Both	81.45	88.07	67.44	77.75	0.8658

Table 5: Accuracies on the VizWiz validation set using information (for both training and inference) from different modalities. In general, questions are difficult to be recognized as unanswerable if given information from only one modality. Thus, both the question and the image are required for a VQA question in VizWiz to be recognized as unanswerable.

Amount of VizWiz Train Dataset Used	Examples Used for Fine-tuning	Accuracy	Answerable Questions Accuracy	Unanswerable Questions Accuracy	Balanced Accuracy	Answerable Questions F1 Score
0%	0	40.77	39.84	42.74	41.29	0.4775
1%	185	71.24	97.10	16.46	56.78	0.8210
5%	925	75.85	86.74	52.78	69.76	0.8299
10%	1851	76.11	87.29	52.42	69.85	0.8323
25%	4627	79.00	89.09	57.62	73.36	0.8522
50%	9254	79.93	87.39	64.12	75.75	0.8554
75%	13882	81.52	89.20	65.27	77.23	0.8677
100%	18509	81.45	88.07	67.44	77.75	0.8658

Table 6: Accuracies on the VizWiz validation set after fine-tuning on some subset of the VizWiz training set. Generally, more examples imply greater variance for improvements on the validation set.

Model	Accuracy	Answerable Questions Accuracy	Unanswerable Questions Accuracy	Balanced Accuracy	Answerable Questions F1 Score
Base UNITER	81.45	88.07	67.44	77.75	0.8658
Large UNITER	81.57	88.41	67.08	77.74	0.8670

Table 7: Accuracies on the VizWiz validation set from different UNITER models. The accuracies are very similar for both models. Note that the large model is fine-tuned for 5000 steps as opposed to the base model fine-tuned for 6000 steps.

Amount of VQA Train Dataset Used	Examples From VQA	Accuracy	Answerable Questions Accuracy	Unanswerable Questions Accuracy	Balanced Accuracy	Answerable Questions F1 Score
0%	0	81.01	88.31	65.56	76.93	0.8634
1%	2483	81.18	87.66	67.44	77.55	0.8635
5%	12417	80.76	87.15	67.22	77.19	0.8602
10%	24835	80.88	88.04	65.70	76.87	0.8622
25%	62087	80.53	87.01	66.79	76.90	0.8586
50%	124174	80.27	86.64	66.79	76.71	0.8565
75%	186262	79.95	84.66	69.96	77.31	0.8516
100%	248349	79.86	85.82	67.22	76.52	0.8527

Table 8: Accuracies on the VizWiz validation set after fine-tuning on the full VizWiz training set with some subset of the VQA training set. Accuracy decreased as more VQA data are added.

Amount of VQA Train Dataset Used	Examples From VQA	Accuracy	Answerable Questions Accuracy	Unanswerable Questions Accuracy	Balanced Accuracy	Answerable Questions F1 Score
0%	0	82.67	95.93	62.20	79.06	0.8705
1%	2483	88.92	90.44	86.57	88.51	0.9083
5%	12417	90.43	92.54	87.16	89.85	0.9215
10%	24835	91.00	93.15	87.67	90.41	0.9262
25%	62087	91.51	93.22	88.87	91.05	0.9303
50%	124174	91.97	93.60	89.46	91.53	0.9340
75%	186262	91.95	93.33	89.82	91.58	0.9337
100%	248349	91.99	92.97	90.47	91.72	0.9337

Table 9: Accuracies on the VQA validation set after fine-tuning on the full VizWiz training set with some subset of the VQA training set. Accuracy increased as more VQA data are added. This increase may be due to the overwhelming amount of examples in the VQA dataset as compared to the 18509 examples in the VizWiz dataset. Thus, the model may become more biased towards information in the VQA dataset.

E Training on VizWiz with QRPE

Tables 10 and 11 show results on models that used a subset of both VizWiz and QRPE. Tables 12 and 13 show results on models that are trained on the full VizWiz train set and a subset of QRPE. All models presented, except for the large models in 14, are fine-tuned for 6000 steps.

Amount of Train Dataset Used	Examples From VizWiz	Examples From QRPE	Accuracy	Answerable Questions Accuracy	Unanswerable Questions Accuracy	Balanced Accuracy	Answerable Questions F1 Score
0%	0	0	33.53	3.20	97.76	50.48	0.0615
1%	185	710	72.22	94.21	25.63	59.92	0.8216
5%	925	3549	75.16	89.54	44.69	67.11	0.8304
10%	1851	7097	77.15	88.24	53.65	70.94	0.8399
25%	4627	17743	76.94	85.21	59.42	72.32	0.8339
50%	9254	35486	78.65	85.41	64.33	74.87	0.8446
75%	13882	53229	80.06	87.49	64.33	75.91	0.8564
100%	18509	70972	80.67	87.05	67.15	77.10	0.8595

Table 10: Accuracies on the VizWiz validation set after fine-tuning on a subset of both the VizWiz training set and the QRPE dataset. In general, accuracies increased as more data is added.

F VILLA Training on VizWiz

Tables 15 and 16 show results on models that used VILLA training on the VizWiz train set. All models presented are fine-tuned for 6000 steps.

G Qualitative Examples

Figures 2 contains some correctly labeled image-question pairs. Figures 1 contains an incorrectly labeled answerable image-question pair.

Amount of Train Dataset Used	Examples From VizWiz	Examples From QRPE	Accuracy	Answerable Questions Accuracy	Unanswerable Questions Accuracy	Balanced Accuracy	Answerable Questions F1 Score
0%	0	0	52.89	5.87	99.91	52.89	0.1107
1%	185	710	82.85	90.93	74.77	82.85	0.8413
5%	925	3549	87.85	90.42	85.28	87.85	0.8815
10%	1851	7097	89.17	92.69	85.65	89.17	0.8954
25%	4627	17743	90.69	92.64	88.76	90.70	0.9087
50%	9254	35486	91.74	93.70	89.78	91.74	0.9190
75%	13882	53229	91.98	94.19	89.77	91.98	0.9215
100%	18509	70972	92.08	94.47	89.68	92.08	0.9226

Table 11: Accuracies on the QRPE validation set after fine-tuning on a subset of both the VizWiz training set and the QRPE dataset. In general, accuracies increased as more data is added.

Amount of QRPE Train Dataset Used	Examples From QRPE	Accuracy	Answerable Questions Accuracy	Unanswerable Questions Accuracy	Balanced Accuracy	Answerable Questions F1 Score
0%	0	80.78	89.54	62.24	75.89	0.8636
1%	710	80.62	86.54	68.09	77.31	0.8585
5%	3549	80.81	88.34	64.84	76.59	0.8621
10%	7097	81.34	87.63	68.01	77.82	0.8645
25%	17743	81.06	88.24	65.85	77.04	0.8636
50%	35486	80.99	88.21	65.70	76.96	0.8631
75%	53229	80.39	86.61	67.22	76.91	0.8571
100%	70972	80.67	87.05	67.15	77.10	0.8595

Table 12: Accuracies on the VizWiz validation set after fine-tuning on the full VizWiz training set and a subset of the QRPE dataset. Note that incorporating QRPE does not seem to affect performance as aggressively as VQA.

Amount of QRPE Train Dataset Used	Examples From QRPE	Accuracy	Answerable Questions Accuracy	Unanswerable Questions Accuracy	Balanced Accuracy	Answerable Questions F1 Score
0%	0	68.82	97.08	40.55	68.82	0.7569
1%	710	84.66	90.00	79.32	84.66	0.8543
5%	3549	88.41	91.44	85.37	88.41	0.8875
10%	7097	89.57	92.59	86.55	89.57	0.8988
25%	17743	90.94	93.27	88.62	90.94	0.9115
50%	35486	91.62	93.84	89.40	91.62	0.9180
75%	53229	91.78	94.13	89.42	91.78	0.9197
100%	70972	92.08	94.47	89.68	92.08	0.9226

Table 13: Accuracies on the QRPE validation set after fine-tuning on the full VizWiz training set and a subset of the QRPE dataset. In general, accuracies increased as more QRPE data is added.

Model	Validation Dataset	Accuracy	Answerable Questions Accuracy	Unanswerable Questions Accuracy	Balanced Accuracy	Answerable Questions F1 Score
Base UNITER	VizWiz	80.67	87.05	67.15	77.10	0.8595
Base UNITER	QRPE	92.08	94.47	89.68	92.08	0.9226
Large UNITER	VizWiz	80.13	87.12	65.34	76.23	0.8563
Large UNITER	QRPE	93.28	95.48	91.08	93.28	0.9342

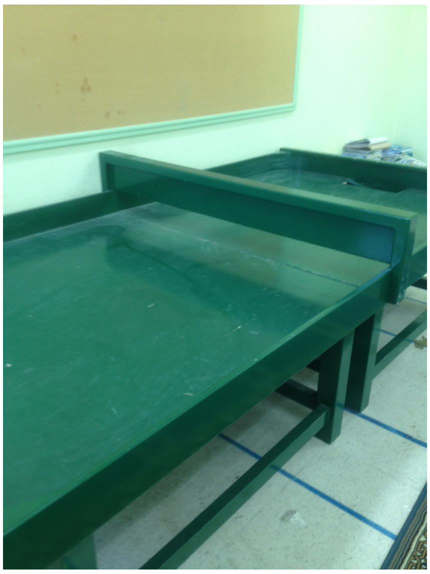
Table 14: Accuracies on the validation sets after fine-tuning on both full VizWiz and QRPE train datasets. Surprisingly, the large UNITER model performed as similar as the base UNITER model on the VizWiz validation set. However, the large UNITER model can do better than the base UNITER model on the QRPE dataset. Note that the large UNITER model is fine-tuned for 5000 steps.

Model	Accuracy	Answerable Questions Accuracy	Unanswerable Questions Accuracy	Balanced Accuracy	Answerable Questions F1 Score
No Perturbations	81.45	88.07	67.44	77.75	0.8658
Random Perturbations of $U(-0.01, 0.01)$	80.83	89.26	62.96	76.11	0.8635
Random Perturbations of $U(-0.001, 0.001)$	80.60	89.06	62.67	75.87	0.8618
Random Perturbations of $U(-0.0001, 0.0001)$	80.85	89.16	63.25	76.21	0.8635
Random Perturbations of $U(-0.00001, 0.00001)$	80.48	88.38	63.75	76.07	0.8602

Table 15: Accuracies on the VizWiz validation set after fine-tuning on the full VizWiz training set using VILLA's training objective but with random perturbations in the text modality only. Interestingly, this training method seem improve answerable questions' accuracy but decrease unanswerable questions' accuracy consistently.

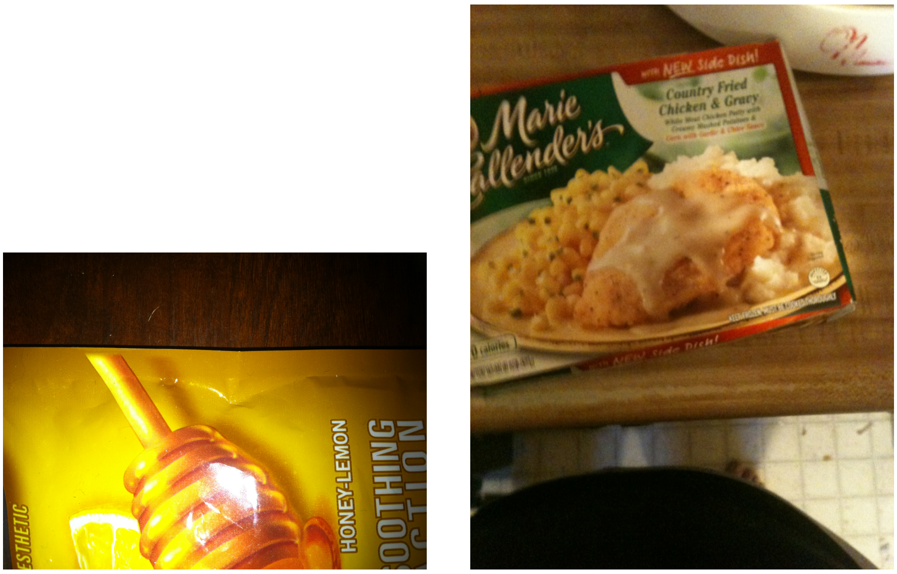
Model	Accuracy	Answerable Questions Accuracy	Unanswerable Questions Accuracy	Balanced Accuracy	Answerable Questions F1 Score
No Perturbations	81.45	88.07	67.44	77.75	0.8658
Random Perturbations of $U(-0.001, 0.001)$ on both modalities	81.66	89.98	64.04	77.01	0.8696
Gradient Updated Perturbations on both modalities	80.92	88.14	65.63	76.89	0.8626
Gradient Updated Perturbations on alternating modalities	81.59	88.95	66.00	77.47	0.8678

Table 16: Accuracies on the VizWiz validation set after fine-tuning on the full VizWiz training set using VILLA's training objective but with different kind of perturbations. Here, "on both modalities" means perturbations are added to the both modalities at a time rather than alternating between them. Consequently, "on alternating modalities" means perturbations are added to the input features of one modality at a time. It seems that using these perturbations seem to improve answerable questions' accuracy but decrease unanswerable questions' accuracy, and overall, improve the accuracy of the model by a tiny bit.



What is this exactly?

Figure 1: Incorrectly labeled as unanswerable when it is answerable.



What is this? Can you tell what dinner this is?

Figure 2: Correctly labeled as answerable questions.