Multi-Modal Contrastive Training for Robust VQA

Abstract

This paper addresses the challenge of enhancing the robustness and efficiency of Visual Question Answering (VQA) models by leveraging feature consistency. Inspired by semi-supervised feature representation learning, we introduce a contrastive loss framework to effectively capture representations from multi-modal inputs. However, existing contrastive learning approaches, which use random intra-class and non-target samples as positive and negative examples, often fail to improve model performance on robust VQA benchmarks. To overcome this limitation, we propose Adversarial Contrastive Learning (ADVCL), a supervised framework that generates challenging positive and negative samples via adversarial perturbations. ADVCL creates hard positives by applying significant perturbations to input image-question pairs, thereby maximizing conditional likelihood and enhancing robustness. Experimental results demonstrate that ADVCL outperforms or matches state-of-the-art models in robustness against linguistic variations in questions, offering a significant advancement in VQA robustness.

Keywords: VQA robustness, contrastive loss, adversarial perturbations

Introduction

Visual Question Answering (VQA) (Antol et al. 2015) is a key application in multi-modal learning that aims to provide accurate answers to textual questions based on image input. Despite significant progress, many VQA models fail to deliver consistent predictions for semantically similar questions posed in different ways (e.g., paraphrased questions). For instance, a model might provide inconsistent answers to "How many cats are in this picture?" and "What is the total number of cats?" despite their semantic equivalence. This issue arises from overreliance on biases in question types and limited attention to holistic question-image semantics.

Recent efforts to address these challenges have introduced robust models that mitigate language bias, often through data augmentation techniques that generate question paraphrases ((Shah et al. 2019; Gokhale et al. 2020; Liang et al. 2020b; Kant et al. 2020; Ghosh and Lan 2021)). While effective to some extent, these methods often fail to fully address linguistic variations, leaving room for improvement in robustness. Moreover, contrastive learning, which has shown promise in single-modal tasks (van den Oord, Li, and Vinyals 2018; Chen et al. 2020b), remains underexplored in multi-modal contexts for VQA.

In this paper, we propose Adversarial Contrastive Learning (ADVCL), a framework designed to improve VQA robustness through a novel application of supervised contrastive loss (Khosla et al. 2020). ADVCL generates challenging positive and negative samples via adversarial perturbations, targeting both visual and textual modalities. By leveraging these adversarial examples, our approach enhances the model's sensitivity to input variations, ensuring more consistent predictions across paraphrased questions.

We evaluate ADVCL on the VQA-Rephrasings benchmark, which tests robustness to linguistic variations, and the widely-used VQA v2.0 dataset (Goyal et al. 2017). ADVCL outperforms state-of-the-art models on VQA-Rephrasings, achieving a 1.57% improvement in consensus score (CS4) over the baseline, and improves overall VQA accuracy on VQA v2.0 by 0.78%. The main contributions include the following:

- We address the VQA robustness problem via adversarial contrastive learning, enhancing representation learning for both vision and language inputs.
- We propose a novel method for generating adversarial positive and negative samples to optimize contrastive loss.
- We introduce distinct generator modules for visual and textual modalities, ensuring robustness to manipulations in both.

Related Works

Visual Question Answering

Visual Question Answering (VQA) was initially introduced to gain a deep understanding of visual content by combining advancements in natural language processing (NLP) and computer vision (CV) (Malinowski and Fritz 2014)). Early methods used pre-trained visual feature extractors like VGG and ResNet, while modern approaches employ more advanced architectures, including multi-modal transformers (Lu et al. 2019, 2020). Notable advancements include the application of bilinear attention (Kim et al. 2016) and

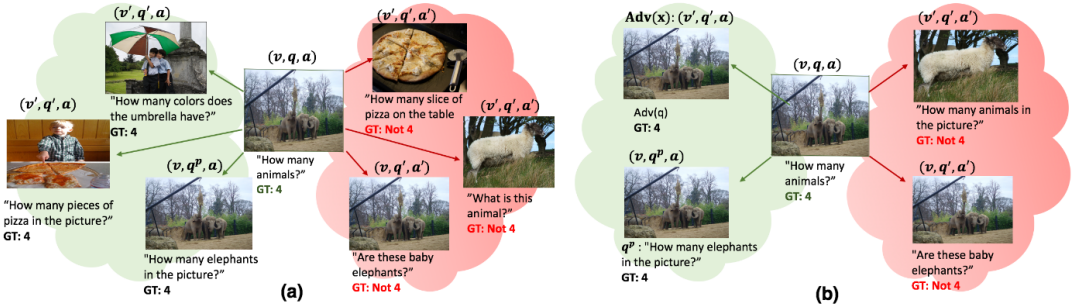

Figure 1: **Overview of proposed adversarial contrastive learning (b) VS. contrastive learning proposed by (Kant et al. 2020) (a).** Our model alleviates the biases from feature by ignoring non-sense intra-class but adding adversarial sample

methods such as BAN (Kim, Jun, and Zhang 2018), and DCN (Nguyen and Okatani 2018).

Recent models like Flamingo (Alayrac et al. 2022) have introduced cross-modal few-shot learning capabilities, enabling robust VQA with limited task-specific examples. Another recent addition is BLIP (Li et al. 2022), which integrates vision-language pretraining for zero-shot VQA across diverse datasets. Datasets like GQA (Hudson and Manning 2019) focus on improving visual grounding and compositional reasoning in VQA. Despite advancements, robust prediction against input variations remains challenging.

Robustness of Visual Question Answering

The robustness of VQA models has been extensively studied with respect to biases in multimodal datasets. (Agrawal et al. 2018) introduced VQA-CP to address question-oriented language bias, while (Shah et al. 2019) highlighted linguistic vulnerabilities using the VQA-Rephrasings dataset. Methods like LMH (cla 2019) and CSS (Chen et al. 2020a) employ debiasing strategies to reduce the impact of spurious correlations in training data.

Recent works emphasize broader robustness across modalities and reasoning. ViLT (Kim, Cho, and Bansal 2021) and BEiT-3 (Wang et al. 2023) extended pre-training paradigms to better align visual and textual representations, showing improvements in generalization.

In addition to tackle linguistic bias, recent benchmarks like GQA-OOD (Kervadec et al. 2020) and CLEVR-Ref+ (Liu et al. 2019) push models to address robustness in compositional and relational reasoning. VL-BERT (Su et al. 2020) and UNITER (Chen et al. 2019) use shared embeddings to enhance robustness against noisy data and adversarial inputs.

Contrastive Representation Learning

Contrastive learning has demonstrated strong performance in learning high-level visual and textual representations (van den Oord, Li, and Vinyals 2018). Early works focused on unsupervised methods, while supervised contrastive learning methods like SimCLR (Chen et al. 2020b) and MoCo (He et al. 2020) improved robustness by aligning semantically similar features.

Recent VQA applications of contrastive loss include using Debiasing Contrastive Learning (DCL) (Jiang et al. 2022) to mitigate dataset-induced biases. (Liang et al. 2020a) explored robust feature learning by replacing cross-entropy with supervised contrastive loss, improving generalization. DeVLBert (Parmar, Jaiswal, and Sharma 2023) extends this idea by incorporating cross-modal contrastive objectives to enhance multi-modal feature alignment. In this work, we focus on adversarially generated positive and negative samples to improve model consistency against both linguistic and visual perturbations, a direction less explored in the existing literature.

Approach

We introduce AdvCL (Adversarial Contrastive Learning), an extension of contrastive learning designed to enhance the robustness of Visual Question Answering (VQA) models against linguistic variations in questions. This approach augments training data and optimizes multi-modal representations by leveraging adversarially generated examples and supervised contrastive learning.

Dataset Augmentation with Question Paraphrases

To enrich the dataset $D = \{(v_i, q_i, a_i)\}_{i=1}^{N}$ comprising triplets of images v_i, question q_i and ground-truth answer a_i, we augment D with paraphrased questions $Q^{Para}(q)$. These paraphrases are generated using two complementary methods: Visual Question Generation (VQG) (Shah et al. 2019) and Back-Translation (BT) (Edunov et al. 2018). The augmented dataset $D^{aug} = D \cup Q^{Para}(q)$.

Contrastive Learning Framework for VQA

Building on prior works (van den Oord, Li, and Vinyals 2018; Chen et al. 2020b), contrastive learning aims to align semantically similar representations while pushing apart dissimilar ones. For VQA, previous methods (Liang et al. 2020b; Kant et al. 2020; Ghosh and Lan 2021) utilize this framework by maximizing mutual information between original and paraphrased samples (positives) while distancing non-target samples (negatives). ADVCL refines this strategy by introducing adversarial positives and negatives to address the limitations of random sampling.

Adversarial Contrastive Learning (ADVCL)

Traditional contrastive learning methods rely on randomly sampling positive and negative pairs, which may lead to suboptimal performance due to the random nature of negative samples. To improve this, we propose Adversarial Contrastive Learning (ADVCL), which refines the sampling process by strategically selecting hard positives and hard negatives using adversarial perturbations.

In ADVCL, we use ground-truth answer labels to generate challenging adversarial examples by perturbing the image and question pairs. These perturbations maximize the loss while maintaining semantic consistency with the original pair, ensuring they are distinct in visual or textual attributes. The perturbation is designed to refine the training by producing more informative samples, enhancing model robustness. For a given triplet (v_i, q_i, a_i), where V_i, q_i, and a_i is the ground-truth answer, we perturb the image and question to create adversarial samples v^{adv} and q^{adv}. These perturbations are designed to maximize the loss function, ensuring the adversarial examples maintain their relationship with the original ground-truth answer a_i. The perturbations are computed as follows:

$$q_{adv} = q + \delta_q^*, \;\; where\; \delta_q^* = argmax\mathcal{L}(q + \delta_q)$$
$$\mathrm{v}_{adv} = v + \delta_v^*, \;\; where\; \delta_v^* = argmax\mathcal{L}(v + \delta_v) \quad (1)$$

The adversarial contrastive loss is calculated using the adversarial pairs created by perturbing the image and question. Hard negatives are generated by altering either the visual or textual components while maintaining semantic similarity in the embedding space. These hard negatives, along with the adversarial pairs, improve the model's ability to distinguish between similar and dissimilar samples, enhancing its robustness across varying inputs. By minimizing the contrastive loss for these pairs, the model learns to create semantically rich representations that generalize better across different question and image variations.

The loss formulation uses a supervised contrastive loss ($\mathcal{L}_{SCL}$ (Khosla et al. 2020)) to pull together positive samples and push apart negative ones:

$$\mathcal{L}_{SCL} = \frac{1}{|P|} \cdot \sum_{j=1}^{P} log \frac{e^{(\phi(z_i.zj)/\tau)}}{\sum_{k=1}^{K} I_{k \neq i}.e^{(\phi(z_i.z_k)/\tau)}} \quad (2)$$

where P denotes the set of positives, z_i and z_j are the embeddings of similar samples, and $\phi(z_i, z_j)$ computes similarity between them (e.g., cosine similarity). Temperature $\tau \neq 0$ controls the smoothness of the similarity function.

In ADVCL, adversarial perturbations are applied to both visual and textual inputs, generating adversarial examples to improve model robustness. From the augmented data set Q^{Para}, four positive samples are created for each input question q_i as ${q_{para}}_p{}_{p=1}^{4}$. ADVCL further generates visual and textual adversarial examples on-the-fly, obtaining semantically equivalent samples for both modalities. Specifically, for each triplet (v_i, q_i, a_i), adversarial pairs are formed as:

$$ADV(x) = (v_a dv, q_i), (v_i, q_a dv), (v_a dv, q_a dv) \quad (3)$$

These pairs, as shown in Figure 1.b, maintain the same ground-truth answer and are generated from the same triplet in the embedding space. To create these adversarial examples, we apply the supervised contrastive loss (LSCL) as the primary loss function, with additional regularization terms that weight the contributions of adversarial samples. The overall loss function is given by:

$$\mathcal{L} = \mathcal{L}_{SCL}(\theta;\; v;\; q;\; a) + \beta_1 \mathcal{L}_{SCL}(\theta;\; v;\; q_{adv};\; a)$$
$$+ \beta_2 \mathcal{L}_{SCL}(\theta;\; v_{adv};\; q;\; a) \quad (4)$$

θ, β_1, and β_2 are hyper-parameters controlling the relative weight of the adversarial samples, as discussed in (Tang et al. 2020). The adversarial samples are generated using the Iterative Fast Gradient Sign Method (IFGSM) (Kurakin, Goodfellow, and Bengio 2017), an efficient gradient-based attacker, which iteratively perturbs the question and image input based on the gradients of the loss function:

$$q_{adv}^{t+1} = q_{adv}^{t} + \alpha \cdot sign(\nabla_q \mathcal{L}(\theta; v; q_{adv}^{t}; a_{true})),$$
$$\mathrm{v}_{adv}^{t+1} = v_{adv}^{t} + \alpha \cdot sign(\nabla_v \mathcal{L}(\theta; v_{adv}^{t}; q; a_{true})). \quad (5)$$

For first step ($t = 1$): $q_{adv}^{1} = q + \alpha.sign(\bigtriangledown_q \mathcal{L}(\theta; v; q; a))$, and $v_{adv}^{1} = v + \alpha.sign(\bigtriangledown_v \mathcal{L}(\theta; v; q;\; a))$. To further enhance model performance, ADVCL avoids the use of random non-target samples as negatives, which is common in traditional contrastive learning frameworks. Instead, it selects two types of hard negatives: (i) samples with similar questions but different answer labels, and (ii) samples with similar visual inputs but different questions and answers, forming pairs $x' = \{(v', q, a'), (v, q', a') \mid a \neq a'\}$. These hard negatives refine the model's ability to differentiate between samples, improving robustness across varying inputs and helping the model generalize better on challenging VQA benchmarks.

Learning ADVCL Overall Loss

The proposed VQA model combines contrastive and cross-entropy training. In this case, the final softmax classifier is learned by minimizing joint loss $\mathcal{L}$ with cross-entropy loss; $\mathcal{L}_{CE}$ and supervised contrastive loss $\mathcal{L}_{SCL}$, which is formulated as $\mathcal{L} = \mathcal{L}_{CE} + \mathcal{L}_{SCL}$.

Experiments

We evaluated the ability of the proposed method to learn robust representations with consistency and discrimination.

Implementation Details

Hard positive and negative samples were generated to enable contrastive learning, emphasizing meaningful input variations. Models were trained using AdamX optimizer (Kingma and Ba 2015), with learning rate of 1e-4 and $\beta = 0.98$. Experiments have been conducted on 3 GTX 1080TI GPU with 60 and 128 batch sizes for contrastive and cross-entropy learning due to the limitations of memory. The learning rate lr and β are both initialized with 0.1. ResNet ((He et al. 2016)) backbone is used for all models except for the black-box experiments.

Model	CS				VQA Scores	
	$k=1$	$k=2$	$k=3$	$k=4$	Orig	Rep
MUTAN (Gokhale et al. 2020)	56.7	43.6	38.9	32.7	59.1	46.8
BUTD (Anderson et al. 2018)	60.5	46.9	40.4	34.5	61.5	51.2
BUTD+CC (Shah et al. 2019)	61.7	50.8	44.7	42.5	62.4	52.6
Pythia (Jiang et al. 2018)	63.4	52.0	45.9	39.5	64.1	54.2
Pythia+CC (Shah et al. 2019)	64.4	55.4	50.9	44.3	64.5	55.6
BAN (Kim, Jun, and Zhang 2018)	64.8	53.1	47.4	39.9	65.0	55.8
BAN+CC (Shah et al. 2019)	65.7	56.9	51.7	48.2	65.8	56.6
ConClaT (Kant et al. 2020)	-	-	55.3	52.3	-	64.7
Ours-ADV	66.8	59.5	55.1	51.9	-	66.0
Ours-ADVCL	**67.1**	**59.7**	**55.3**	**52.2**	**67.3**	**70.1**

Table 1: Consensus performance on VQA-Rephrasings dataset using VQG Baseline results are copied from (Shah et al. 2019)

Comparison with State-of-the-Art

We compared ADVCL with classic: PYTHIA (Jiang et al. 2018), Bottom-Up-Attention and Top-Down (BUTD) (Anderson et al. 2018) and state-of-the-art settings: VQA+CC (Shah et al. 2019) and Contrast and Classify Training (CONCLAT) (Kant et al. 2020). Besides, we used two baselines: first, the Multi-Modal Transformer-based model (MMT) using CE loss ($\mathcal{L}_{CE}$), that is denoted as BASELINE. The second, denoted as Ours-ADV is MMT with adversarial attack training that uses $\mathcal{L}_{CE}$ for both clean and noisy status. Ours-ADV is very close to the model proposed by VILLA ((Gan et al. 2020)). For fair comparison, we have searched for methods using the same Faster-RCNN features ((Ren et al. 2017)) similar to ours. We use the evaluation code from official VQA challenge ((Antol et al. 2015)).

Table 1 reports the comparison of our model performance with various state-of-the-art methods on the consensus score (CS(k)) for $k = 1, 2, 3, 4$ on VQA-Rephrasings (Shah et al. 2019) and VQA Accuracy on VQA v2.0 ((Goyal et al. 2017)) datasets. For fair comparison, we provide CS(k) performances on augmented data by VQG. Our method outperforms Conclat gains of 1.2%, on validation (CS(k)) for $k = 3$ and $k = 4$ respectively. Table 1 is also provides further comparison between our proposed model and state-of-the-arts due to original dataset and augmented dataset using positive question rephrasing (Rep). In summary, OURS-ADVCL achieves state-of-the-art robust performance on the VQA-Rephrasings dataset (Shah et al. 2019) by using robust metrics that show the robustness of the proposed model across language variations.

Qualitative Analysis

Table 2 provides the performance based on different question types of some other state-of-the-art methods ((Anderson et al. 2018; Chen et al. 2020a; Cadene et al. 2019; Shah et al. 2019)) to further evaluate the discriminative power of the representations for answer prediction. The results show the promising improvements on both VQA v2.0 and out-of-distribution dataset like VQA-CP v2.0(Agrawal et al. 2018).

Qualitative Analysis

We present qualitative results in Figure 2, comparing the performance of the baseline and ADVCL on various aspects of VQA robustness, including the handling of complex questions (upper row) and biased samples (lower row). For instance, the ground-truth answer to the question "what color is the hydrant?" is frequently "red" due to dataset imbalances. However, for the question "What color is the top of the hydrant?" the correct answer is "white", as shown in Figure 2. This demonstrates the model's ability to distinguish between "top of hydrant" and "hydrant" despite language bias, showcasing the importance of high-level representations to mitigate such biases in visual question answering.

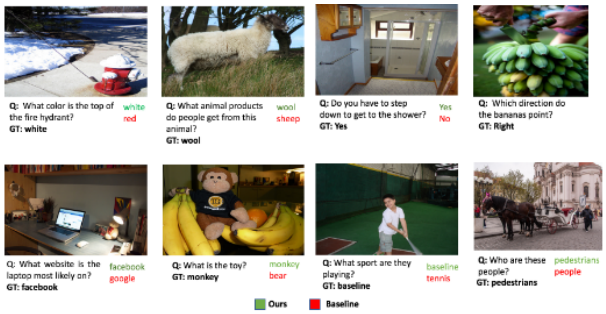

Figure 2: **Qualitative Examples.** Visualization of examples collected from ADVCL predictor for complicated questions and unbiased samples in compare with Ours-ADV.

Additionally, Figure 3 illustrates the ability of ADVCL to learn more consistent representations compared to the baseline. The qualitative results demonstrate improved consistency in predictions across different rephrasings of the original question (**Q1**, **Q2**, **Q3**). These results were generated using data augmented via back-translation (BT).

Conclusion

This paper highlights the significance of learning stable features to enhance the robustness of VQA models against linguistic variations in questions. Specifically, ADVCL improves both model consistency (robustness) and correctness (discrimination) by combining contrastive and cross-entropy loss frameworks. Our method achieves notable improvements on benchmarks. On the VQA-Rephrasings dataset, ADVCL boosts the consensus score (CS@K) by 1.57% over the baseline and surpasses the state-of-the-art score, improving from 48.2 to 53.3. Additionally, on the standard VQA v2.0 benchmark, ADVCL achieves an overall accuracy gain of 0.78%. These results demonstrate the effectiveness of ADVCL in promoting both consistency and robustness in VQA tasks.

References

2019. Don't Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases.

Agrawal, A.; Batra, D.; Parikh, D.; and Kembhavi, A. 2018. Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4971–4980.

Alayrac, J.-B.; Donahue, J.; Elhoseiny, M.; et al. 2022. Flamingo: a Visual Language Model for Few-Shot Learn-

Model	VQA-CP v2.0 test(%)				VQA v2.0 val (%)			
	Yes/No	Num.	Other	Overall	Yes/No	Num.	Other	Overall
GVQA (Agrawal et al. 2018)	-	-	-	31.3	72.0	31.2	34.6	48.2
BUTD (Anderson et al. 2018)	42.3	11.9	46.0	39.7	81.2	42.1	55.6	63.5
RUBi (Cadene et al. 2019)	42.8	12.8	43.2	38.5	-	-	-	63.1
MUREL (Cadène et al. 2019)	42.9	13.2	45.0	39.5	-	-	-	65.1
LXMERT (Tan and Bansal 2019)	42.8	18.9	55.5	46.2	83.3	46.2	56.9	65.3
methods based on data-augmentation and training strategy								
CSS (Chen et al. 2020a)	84.4	49.4	48.2	58.9	73.3	39.8	55.1	59.9
CL-VQA (Liang et al. 2020b)	86.9	49.9	47.2	59.2	67.3	38.4	54.7	57.3
Loss-Rescaling (Guo et al. 2020)	72.8	48.0	44.5	53.3	68.2	36.4	52.3	56.8
MUTANT (Gokhale et al. 2020)	88.9	49.7	50.8	61.7	82.1	42.5	53.3	62.6
RandImg (Teney et al. 2020)	83.9	41.6	44.2	55.4	76.5	33.9	48.6	57.2
Unshuffling (Teney, Abbasnejad, and van den Hengel 2020)	47.7	14.4	47.2	42.4	78.3	42.2	52.8	61.1
ADA-VQA (Guo et al. 2021)	87.4	53.0	46.8	59.6	78.8	42.2	54.4	61.9
Ours-ADV	-	-	-	-	83.0	48.8	57.3	66.0
Ours-ADVCL	**88.3**	**56.0**	**62.8**	**62.1**	**85.2**	**51.2**	**60.1**	**67.3**

Table 2: Comparison to task-specific state-of-the-arts on VQA-CP v2.0 test VQA v2.0 validation split.

Figure 3: **Qualitative Examples.** Predicate of ADVC and our baseline on several image-question pairs and their corresponding rephrased questions.

ing. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*. PMLR.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; and Parikh, D. 2015. Vqa: Visual question answering. In *ICCV*, 2425–2433.

Cadène, R.; Ben-younes, H.; Cord, M.; and Thome, N. 2019. MUREL: Multimodal Relational Reasoning for Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Cadene, R.; Dancette, C.; Ben-younes, H.; Cord, M.; and Parikh, D. 2019. RUBi: Reducing Unimodal Biases in Visual Question Answering. In *arXiv preprint arXiv:1906.10169*.

Chen, L.; Yan, X.; Xiao, J.; Zhang, H.; Pu; and Zhuang. 2020a. Counterfactual Samples Synthesizing for Robust Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, T.; Kornblith; Norouzi, M.; and Hinton, G. E. 2020b. A Simple Framework for Contrastive Learning of Visual Representations. In *ICML*.

Chen, Y.; Li, L.; Yu, L.; Kholy, A. E.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2019. UNITER: Learning UNiversal Image-TExt Representations. *arXiv*.

Edunov, S.; Ott, M.; Auli, M.; and Grangier, D. 2018. Understanding Back-Translation at Scale. 489–500.

Gan, Z.; Chen, Y.; Li, L.; Zhu, C.; Cheng, Y.; and Liu., J. 2020. Large-Scale Adversarial Training for Vision-and-Language Representation Learning. In *NIPS*.

Ghosh, A.; and Lan, A. S. 2021. Contrastive Learning Improves Model Robustness Under Label Noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2703–2708.

Gokhale, T.; Banerjee, P.; Baral, C.; and Yang, Y. 2020. MUTANT: A Training Paradigm for Out-of-Distribution Generalization in Visual Question Answering. In *EMNLP*, 878–892.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. 6904–6913.

Guo, Y.; Nie, L.; Cheng, Z.; Ji, F.; Zhang, J.; and Bimbo, A. D. 2021. AdaVQA: Overcoming Language Priors with Adapted Margin Cosine Loss. In *IJCAI*, 708–714.

Guo, Y.; Nie, L.; Cheng, Z.; and Tian, Q. 2020. Loss-rescaling VQA: Revisiting Language Prior Problem from a Class-imbalance View. *arXiv*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. B. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9726–9735.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Hudson, D. A.; and Manning, C. D. 2019. GQA: a new dataset for compositional question answering over real-world images. *arXiv*.

Jiang, Y.; Li, Y.; Wang, P.; Zhang, Z.; and Liu, X. 2022. DCL: Debiased Contrastive Learning for Visual Representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 782–791.

Jiang, Y.; Natarajan, V.; Chen, X.; Rohrbach, M.; Batra, D.; and Parikh, D. 2018. Pythia v0.1: the Winning Entry to the VQA Challenge 2018. *arXiv preprint arXiv:1807.09956*.

Kant, Y.; Moudgil, A.; Batra, D.; Parikh, D.; and Agrawal, H. 2020. Contrast and Classify: Alternate Training for Robust VQA. In *arXiv*.

Kervadec, H.; Picard, D.; Ben, A.; and Delfosse, V. 2020. Roses are Red, Violets are Blue... But Should VQA expect Them To? *arXiv preprint arXiv:2006.05121*.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised Contrastive Learning. In *NIPS*.

Kim, J.; Jun, J.; and Zhang, B. 2018. Bilinear Attention Networks. In *NIPS*.

Kim, J.; Lee, S.; Kwak, D.; Heo, M.; Kim, J.; Ha, J.; and Zhang, B. 2016. Multimodal Residual Learning for Visual QA. In *Advances in Neural Information Processing Systems*, 361–369.

Kim, W.; Cho, K.; and Bansal, M. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 1–12.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

Kurakin, A.; Goodfellow, I.; and Bengio. 2017. Adversarial examples in the physical world. In *ICLR*.

Li, J.; Li, X.; Lu, Y.; Zhang, Z.; Zhao, L.; Xu, W.; Chen, Y.; Yang, Y.; Zhang, L.; Li, B.; et al. 2022. BLIP: Bootstrapping Language-Image Pretraining with Frozen Image Encoders and Large Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14234–14243.

Liang, X.; He, J.; Zeng, J.; He, X.; Wei, F.; Zhu, X.; Zhang, M.; and Li, S. 2020a. XGLUE: A New Benchmark Dataset for Cross-lingual Pre-training, Understanding, and Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*.

Liang, Z.; Jiang, W.; Hu, H.; and Zhu, J. 2020b. Learning to Contrast the Counterfactual Samples for Robust Visual Question Answering. In *EMNLP*.

Liu, R.; Liu, C.; Bai, Y.; and Yuille, A. L. 2019. CLEVR-Ref+: Diagnosing Visual Reasoning With Referring Expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4185–4194.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. 13–23.

Lu, J.; Goswami, V.; Rohrbach, M.; Parikh, D.; and Lee, S. 2020. 12-in-1: Multi-Task Vision and Language Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Malinowski, M.; and Fritz, M. 2014. A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input. In *NIPS*, 1682–1690.

Nguyen, D.; and Okatani, T. 2018. Improved Fusion of Visual and Language Representations by Dense Symmetric Co-Attention for Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6087–6096.

Parmar, N.; Jaiswal, A.; and Sharma, A. 2023. DeVLBert: Improving Vision-Language Pretraining with Deconfounded Learning for Better Visual-Linguistic Understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6).

Shah, M.; Chen, X.; Rohrbach, M.; and Parikh, D. 2019. Cycle-Consistency for Robust Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6649–6658.

Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; and Dai, J. 2020. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *ICLR*.

Tan, H.; and Bansal, M. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *EMNLP*, 5099–5110.

Tang, R.; Ma, C.; Zhang, W. E.; Wu, Q.; and Yang, X. 2020. Semantic Equivalent Adversarial Data Augmentation for Visual Question Answering. In *European Conference on Computer Vision(ECCV)*, 437–453.

Teney, D.; Abbasnejad, E.; Kafle, K.; Shrestha, R.; Kanan, C.; and van den Hengel, A. 2020. On the Value of Out-of-Distribution Testing: An Example of Goodhart's Law. In *NeurIPS*.

Teney, D.; Abbasnejad, E.; and van den Hengel, A. 2020. Unshuffling Data for Improved Generalization. *arXiv preprint arXiv:2002.11894*.

van den Oord, A.; Li, Y.; and Vinyals, O. 2018. Representation Learning with Contrastive Predictive Coding. *arXiv*.

Wang, X.; Lu, Y.; Zhang, K.; Xu, Z.; Zhang, P.; Li, Y.; Liu, F.; Xie, L.; Zhang, L.; and Lu, Y. 2023. BEiT-3: Vision-Language Pretraining with Image-Text Pairs. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 1–12.