Learning to Contrast the Counterfactual Samples for Robust Visual Question Answering

Zujie Liang, Weitao Jiang, Haifeng Hu, Jiaying Zhu
School of Electronics and Information Technology, Sun Yat-sen University
{liangzj9,jiangwt5,zhujy53}@mail2.sysu.edu.cn, huhaif@mail.sysu.edu.cn

Abstract

In the task of Visual Question Answering (VQA), most state-of-the-art models tend to learn spurious correlations in the training set and achieve poor performance in out-of-distribution test data. Some methods of generating counterfactual samples have been proposed to alleviate this problem. However, the counterfactual samples generated by most previous methods are simply added to the training data for augmentation and are not fully utilized. Therefore, we introduce a novel self-supervised contrastive learning mechanism to learn the relationship between original samples, factual samples and counterfactual samples. With the better cross-modal joint embeddings learned from the auxiliary training objective, the reasoning capability and robustness of the VQA model are boosted significantly. We evaluate the effectiveness of our method by surpassing current state-of-the-art models on the VQA-CP dataset, a diagnostic benchmark for assessing the VQA model's robustness.

1 Introduction

To develop human-like visual and language understanding of AI, the task of answering a question about the given visual content has been proposed, *i.e.*, Visual Question Answering (VQA) (Antol et al., 2015). Although the current state-of-the-art methods (Fukui et al., 2016; Anderson et al., 2018; Cadene et al., 2019a) can achieve good results on the VQA benchmarks such as VQA v2 (Goyal et al., 2017), recent researches (Agrawal et al., 2016; Kafle and Kanan, 2017; Agrawal et al., 2018) have found that these methods tend to explore superficial correlations in the training set and perform poorly when transferred to real world setting. Specifically, given a question "What color is the banana?", the models prefer to take the shortcut and "assume" that the answer should be "yellow" since it is the most common answer in the training set, rather than be grounded on the image. To overcome the language bias problems in VQA, (Agrawal et al., 2018) have proposed a dataset named VQA-CP, where the answer distribution of the training set differs from the test set vastly. The performance of most current state-of-the-art models (Andreas et al., 2016; Teney et al., 2018; Shrestha et al., 2019) drop significantly on the VQA-CP due to the language bias. Hence, it has become the standard out-of-distribution benchmark for VQA.

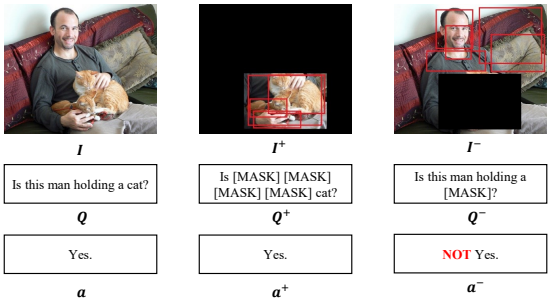

Figure 1: An informal examples of original sample (I, Q), factual sample (I^+, Q^+) and counterfactual sample (I^-, Q^-) generated by the counterfactual sample synthesizing algorithm (Chen et al., 2020).

A successful robust and unbiased VQA system is supposed to be able to deduce the right answer from the right area of the image. Lately, some studies have proposed to synthesize counterfactual samples to improve the robustness of VQA models. (Agarwal et al., 2019; Pan et al., 2019) apply GAN (Goodfellow et al., 2014) to generate images. CSS algorithm proposed by (Chen et al., 2020) generates counterfactual samples by masking the critical objects in images or words in questions, as shown in Figure 1. The critical objects or words can be obtained from CSS as by-products. Nevertheless, the counterfactual samples are simply added to the training data for augmentation, ignoring that the relationship between original samples

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 3285–3292,
November 16–20, 2020.

and counterfactual samples are vital for the reasoning of VQA models. Specifically, the model should be able to learn why the correct answer cannot be inferred after changing the original sample to the counterfactual sample. We posit that modeling the relationship between original samples, factual samples and counterfactual samples can bring more self-supervised signals to improve the reasoning ability of the model.

In order to enable the VQA model to understand the impact of the samples changing from original to counterfactual, we introduce a novel contrastive learning mechanism into the training with counterfactual samples, which is first proposed in the field of learning with counterfactual samples. The auxiliary contrastive training objective model the relationship between original samples, factual samples and counterfactual samples in the cross-modal joint embedding space. With the better cross-modal representations, both the reasoning ability and robustness of the VQA model are improved efficiently.

Overall, the contributions of this paper are as follows:

- We are the first to introduce a self-supervised contrastive learning mechanism for counterfactual samples in VQA. Our method not only helps the VQA model learn the relationship between original samples, factual samples and counterfactual samples but also improves the generalization ability of the model significantly.
- Experiment results show that our method brings significant improvements and achieves state of the art on VQA-CP dataset. Furthermore, the effectiveness of contrastive mechanism in counterfactual sample learning is not limited to the form of contrastive loss.

2 Related Work

2.1 Language Bias in VQA

As the issue of language bias in VQA models is pointed out (Agrawal et al., 2016; Jabri et al., 2016; Goyal et al., 2017), creating a more balanced dataset is a simple way to alleviate it. To this end, the VQA v2 dataset (Goyal et al., 2017) rearranges the sample distribution so that it contains at least one different answer when given a same question and a similar image. Since the statistical bias problem remains, (Agrawal et al., 2018) introduce the VQA-CP dataset where the answer distributions are re-distributed in the training and test splits, making it become the standard benchmark for evaluating the robustness of VQA models.

2.2 Counterfactual Samples for VQA

Recently, employing insights from causal inference (Neuberg, 2003), some researches synthesize counterfactual samples to augment the training of VQA models (Agarwal et al., 2019; Pan et al., 2019; Chen et al., 2020). Similar to our work, (Teney et al., 2020a) have proposed a training objective named Gradient Supervision (GS) to use the relation information between original training samples and additional counterfactual samples. The GS encourages the gradient of the model to align with a "ground truth" gradient, which is the translation from original sample to counterfactual sample in the input space. In contrast, we employ a novel contrastive learning strategy to simultaneously learn the triplet relationship between the original training samples, factual samples and counterfactual samples.

2.3 Contrastive Learning

Contrastive learning techniques have achieved great success in unsupervised learning (Oord et al., 2018; He et al., 2019). The main idea of unsupervised contrastive learning is to maximize the mutual information between the input samples and positive samples so as to learn better representations. Inspired by this, we apply the contrastive mechanism to learn the self-supervision information from counterfactual samples for the first time and improve the robustness of VQA models.

3 Methodology

In this section, we introduce our technical realization. The flowchart of our proposed method is illustrated in Figure 2. Our method consists of three parts: (1) A base VQA model (2) A factual and Counterfactual Samples Synthesizing (CSS) module (3) A Contrastive Learning (CL) objective.

3.1 Baseline VQA Model

We adopt the Bottom-Up Top-Down (UpDn) (Anderson et al., 2018) model into our method, which considers the common formulation of VQA task as a multi-class classification problem. Given a set consisting of N triplets of images $I_i \in \mathcal{I}$, question $Q_i \in \mathcal{Q}$ and answer $a_i \in \mathcal{A}$, we denote as $\mathcal{D} = \{I_i, Q_i, a_i\}_{i=1}^{N}$. The task aims to learn a mapping function $f_{vqa} : \mathcal{I} \times \mathcal{Q} \rightarrow [0,1]^{|\mathcal{A}|}$, producing

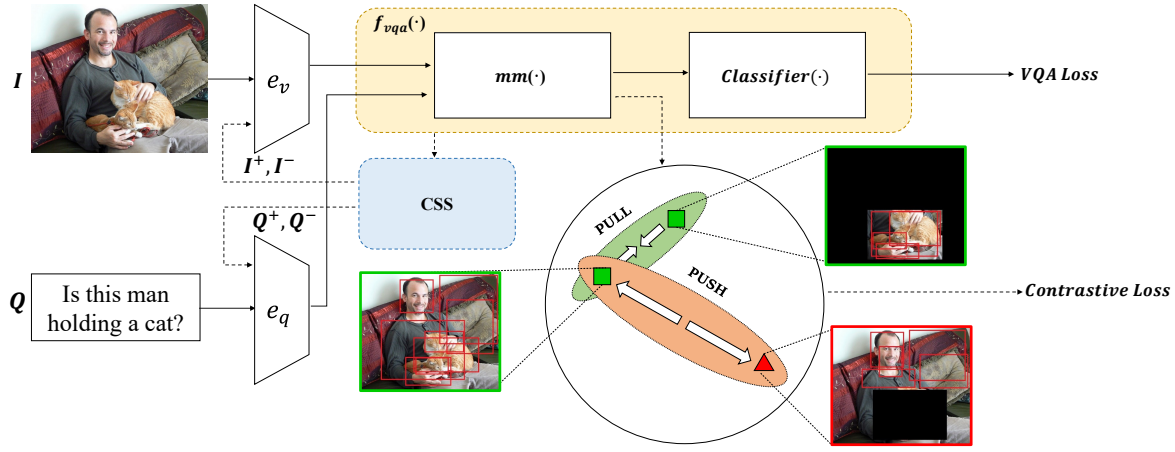

Figure 2: The flowchart of our proposed method. Optimizing the contrastive loss can pull up the original sample (I or Q) and factual sample (I^+ or Q^+) and push away the original sample and counterfactual sample (I^- or Q^-) in the joint embedding space. The example here is the case of (I, I^+, I^-).

an answer distribution of the given image and question. In the following sections, we will omit the subscript i for simplicity. For each question Q, the UpDn uses a question encoder e_q to extract a set of word embeddings $\boldsymbol{Q}$. For each image I, the UpDn uses an object detector e_v to extract a set of visual object embeddings $\boldsymbol{V}$. Then both $\boldsymbol{Q}$ and $\boldsymbol{V}$ are fed into attention and fusion modules to generate the joint embedding $mm(\boldsymbol{Q}, \boldsymbol{V})$. The joint embedding is then fed into classifier C to predict the answer:

$$P_{vqa}(\mathbf{a}|I, Q) = f_{vqa}(\boldsymbol{V}, \boldsymbol{Q}) = C(mm(\boldsymbol{Q}, \boldsymbol{V})) \quad (1)$$

3.2 Synthesizing Counterfactual Samples

There are several ways to synthesize the counterfactual samples of the given image-question pairs in our pipeline. For instance, (Teney et al., 2020a) build counterfactual samples using annotations of human attention (Das et al., 2016). Basically, they generate the counterfactual image by masking the features whose bounding boxes overlap with the human attention map past a certain threshold. In contrast to using extra manual annotations, CSS algorithm proposed by (Chen et al., 2020) calculates the critical objects (I^+) in image or words (Q^+) in question by the modified Grad-CAM (Selvaraju et al., 2017) and masks them to generate the counterfactual samples. Since the latter is more practical, we adopt the CSS algorithm into our pipeline and obtain the factual (I^+, Q^+) and counterfactual (I^-, Q^-) samples:

$$(I^+, I^-, Q^+, Q^-) = CSS(f_{vqa}, (I, Q, a)) \quad (2)$$

3.3 Contrastive Learning Objective

With the causal triplets (I, I^+, I^-) and (Q, Q^+, Q^-) obtained from CSS, we can apply the contrastive learning mechanism. We take a specific triplet (I, I^+, I^-) as an example shown in Figure 2 to illustrate the contrastive learning method. First, the I, I^+ and I^- paired with the Q are fed into the VQA model to generate the joint embeddings of them. Then, we denote the joint embedding $mm(\boldsymbol{Q}, \boldsymbol{V})$ of the original sample as the anchor a, the embedding $mm(\boldsymbol{Q}, \boldsymbol{V}^+)$ of the factual sample as the positive p and the embedding $mm(\boldsymbol{Q}, \boldsymbol{V}^-)$ of the counterfactual sample as the negative n.

Before defining the contrastive loss, we first define a scoring function s that outputs high values for the positive sample and low values for the negative sample. We take the cosine similarity of the representations in the joint embedding space as our scoring function because it implicitly normalizes the embeddings. The score between the anchor and the positive $s(a, p)$ can be described as:

$$s(a, p) = \frac{a^T \cdot p}{\|a\| \cdot \|p\|} \quad (3)$$

Similarly, the score between the anchor and the negative is defined as $s(a, n)$. Then, following recent work in unsupervised learning (Oord et al., 2018), the contrastive loss is formulated as:

$$L_c = \mathop{\mathbb{E}}_{a,p,n}\left[-\log\left(\frac{e^{s(a,p)}}{e^{s(a,p)} + e^{s(a,n)}}\right)\right] \quad (4)$$

For each synthesized triplet, minimizing this loss

Model	Expl.	VQA-CP v2 *test*			
		Overall	Y/N	Number	Other
SAN (Yang et al., 2016)		24.96	38.35	11.14	21.74
GVQA (Agrawal et al., 2018)		31.30	57.99	13.68	22.14
Unshuffling (Teney et al., 2020b)		42.39	47.72	14.43	47.24
+CF (Teney et al., 2020a)	HAT	46.00	61.30	15.60	46.00
+CF+GS (Teney et al., 2020a)	HAT	46.80	64.50	15.30	45.90
UpDn (Anderson et al., 2018)		39.74	42.27	11.93	46.05
+AReg (Ramakrishnan et al., 2018)		41.17	65.49	15.48	35.48
+GRL (Grand and Belinkov, 2019)		42.33	59.74	14.78	40.76
+RUBi (Cadene et al., 2019b)		44.23	67.05	17.48	39.61
+LMH (Clark et al., 2019)		52.01	72.58	31.12	46.97
+LMH+CSS* (Chen et al., 2020)		57.74	83.18	47.59	47.19
+LMH+CSS+GS* (Teney et al., 2020a)		57.37	79.71	**50.85**	47.45
+LMH+CSS+CL(ours)		**59.18**	**86.99**	49.89	47.16
+HINT (Selvaraju et al., 2019)	HAT	47.70	70.04	10.68	46.31
+SCR (Wu and Mooney, 2019)	HAT	49.17	71.55	10.72	**47.49**

Table 1: Performance (%) comparison with SoTA on VQA-CP v2 dataset. *indicates the results of our reimplementation. Expl. denotes the extra annotations that the model has used. HAT is the human attention (Das et al., 2016).

can maximize a lower bound on mutual information between factual sample and original sample, enabling the model to learn the relationship between them and predict the right answer from a more causal aspect. The weighted sum of this contrastive loss and the base VQA classification loss L_{vqa} make up the overall loss:

$$L = \lambda_{vqa} L_{vqa} + \lambda_c L_c \tag{5}$$

where λ_{vqa} and λ_c are the loss weight for each loss.

4 Experiments

4.1 Datasets

The VQA-CP dataset[1] (Agrawal et al., 2018) is the standard benchmark for evaluating the robustness of VQA models, where the answer distribution of the training set differs from the test set vastly. The VQA-CP v1 *train* consists of ∼118K images, ∼245K questions and ∼2.5M answers (∼121K images, ∼438K questions and ∼4.4M answers for VQA-CP v2 *train*). The VQA-CP v1 *test* consists of ∼87K images, ∼125K questions and ∼1.3M answers (∼ 98K images, ∼220K questions and ∼2.2M answers for VQA-CP v2 *test*).

4.2 Settings and Comparisons with SoTA

We validate the effectiveness of our method in the VQA-CP (both v1 and v2) datasets (Agrawal et al., 2018). Results on the VQA v2 are also reported in appendices for completeness. We use the standard VQA evaluation metric (Antol et al., 2015) for accuracy report. All our implementation details are in appendices.

[1]https://www.cc.gatech.edu/ aagrawal307/vqa-cp/

4.3 Performance on VQA-CP v2

Table 1 shows the result comparison with the state-of-the-art models on the VQA-CP v2. According to the backbone of these models, we group them into: 1) **SAN** based methods, including **GVQA**. 2) **Unshuffling** based methods, including **CF**, **CF+GS**. 3) **UpDn** based methods, including **AReg**, **GRL**, **RUBi**, **LMH**, **CSS**, **HINT** and **SCR**. The results show that our Contrastive Learning (CL) building on top of UpDn+LMH+CSS outperforms these previous results, improving the overall accuracy from 57.74% to 59.18% (+1.44%). In contrast, the Gradient Supervision (GS) for the counterfactuals brings smaller gain (+0.80%) from Unshuffling+CF. We further explore the performance of Gradient Supervision when applied with the same set of counterfactual samples (CSS). From Table 1, we can observe that our method still outperforms the LMH+CSS+GS by 1.88%, indicating that our method can bring more self supervision from the counterfactual samples than the GS.

4.3.1 Performance on VQA-CP v1

Table 2 shows performance comparisons with the existing state-of-the-art methods on the VQA-CP v1 test split. We achieves a new state-of-the-art performance on VQA-CP v1 test split, improving the UpDn+LMH+CSS method from 59.63% to 61.27% (+1.64%). Particularly, our method outperforms the Gradient Supervision(GS) by 3.22%.

Model	VQA-CP v1 *test*			
	Overall	Y/N	Number	Other
UpDn (Anderson et al., 2018)	37.87	42.58	14.16	42.71
+AReg (Ramakrishnan et al., 2018)	45.69	77.64	13.21	26.97
+GRL (Grand and Belinkov, 2019)	44.09	75.01	13.40	25.67
+RUBi (Cadene et al., 2019b)	44.81	69.65	14.91	32.13
+LMH (Clark et al., 2019)	55.27	76.47	26.66	45.68
+LMH+CSS* (Chen et al., 2020)	59.63	86.62	28.93	45.12
+LMH+CSS+GS* (Teney et al., 2020a)	58.05	78.50	**37.24**	**46.08**
+LMH+CSS+CL(ours)	**61.27**	**88.14**	34.43	45.34

Table 2: Performance comparison on VQA-CP v1 *test*. *indicates the results of our reimplementation.

4.4 Different Forms of Contrastive Loss

To explore whether different forms of contrastive loss are effective in learning the counterfactual samples in VQA, we conduct experiments on the VQA-CP v2 using the varient of Margin-based Contrastive Loss (MarginCL) proposed by (Hadsell

Model	Overall	Y/N	Number	Other
UpDn (Anderson et al., 2018)	39.74	42.27	11.93	46.05
UpDn*	38.85	42.60	11.51	44.38
+CSS*	39.77	42.80	**12.55**	45.66
+CSS+GS*	40.02	41.97	11.94	46.70
+CSS+MarginCL(ours)	40.15	42.38	12.45	46.57
+CSS+CL(ours)	**40.49**	**42.90**	12.44	**46.93**
LMH (Clark et al., 2019)	52.01	72.58	31.12	46.97
LMH*	52.66	73.47	34.21	46.81
+CSS*	57.74	83.18	47.59	47.19
+CSS+GS*	57.37	79.71	50.85	**47.45**
+CSS+MarginCL(ours)	58.68	85.54	**51.60**	46.54
+CSS+CL(ours)	**59.18**	**86.99**	49.89	47.16

Table 3: Effectiveness of different supervision of counterfactual samples on different architectures on VQA-CP v2 test. *indicates the results of our reimplementation.

et al., 2006), which is formulated as:

$$L_{MC} = D(a,p) + \max(0, m - D(a,n)) \quad (6)$$

where the $D(a,p) = 1 - s(a,p)$ (cosine distance between a and p). The m is the margin between a and n, which is set to 0.3. Table 3 shows the experimental results. The improvements on two different VQA models demonstrate that our method is generic.

4.5 Performance of counterfactual samples and factual samples

To further explore whether our method improves the generalization capability of the VQA model, we conduct the experiments about the VQA performance of the counterfactual samples and factual samples on the VQA-CP v2 and report the result in Table 4. Comparing with the CSS and the CSS+GS, our method achieves the best performance, which demonstrates that the VQA model benefits from the contrastive learning mechanism and accordingly generalizes better on the counterfactual samples and factual samples.

Model	Original Samples	Factual Samples	Counterfactual Samples
CSS	57.74	46.41	48.96
CSS+GS*	57.37	45.83	50.09
CSS+CL(ours)	**59.18**	**46.73**	**50.12**

Table 4: The VQA performance (%) of the counterfactual samples and factual samples on VQA-CP v2 dataset. *indicates the results of our reimplementation.

4.6 Case Study

To validate the effects of our contrastive training objective, we visualize the joint embeddings of two examples and their synthesized samples by employing the t-SNE (Maaten and Hinton, 2008). As Figure 3 shows, compared with the LMH+CSS, our auxiliary training objective helps to not only pull up the original sample and factual sample but also push away the original sample and counterfactual sample in the embedding space, which may build a better causal VQA model.

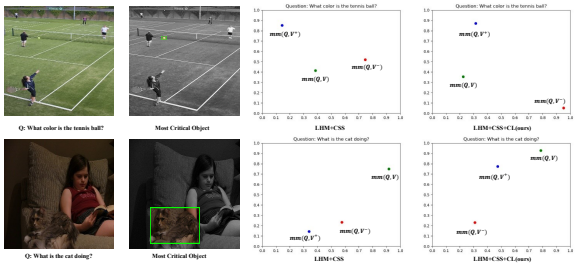

Figure 3: t-SNE visualizations of the cross-modal joint embedding space of the causal triplet generated by the CSS algorithm. $mm(\boldsymbol{Q}, \boldsymbol{V})$ is the joint embedding of original input. $mm(\boldsymbol{Q}, \boldsymbol{V}^+)$ and $mm(\boldsymbol{Q}, \boldsymbol{V}^-)$ are the embeddings of the input with only the most critical object and without the most critical object respectively.

5 Conclusion

In order to fully utilize the supervision information of synthesized counterfactual samples in robust VQA, we introduce a self-supervised contrastive learning mechanism to learn the relationship between factual samples and counterfactual samples. The experimental results demonstrate that our method improves the reasoning ability and robustness of the VQA models.

Acknowledgments

This work has been supported by the National Natural Science Foundation of China under Grants 62076262 and 61673402.

References

Vedika Agarwal, Rakshith Shetty, and Mario Fritz. 2019. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. *arXiv preprint arXiv:1912.07538*.

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the behavior of visual question answering models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1955–1960.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume;

look and answer: Overcoming priors for visual question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *The IEEE International Conference on Computer Vision (ICCV)*.

Remi Cadene, Hedi Ben-younes, Matthieu Cord, and Nicolas Thome. 2019a. Murel: Multimodal relational reasoning for visual question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Remi Cadene, Corentin Dancette, Hedi Ben younes, Matthieu Cord, and Devi Parikh. 2019b. Rubi: Reducing unimodal biases for visual question answering. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 841–852. Curran Associates, Inc.

Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. 2020. Counterfactual samples synthesizing for robust visual question answering. In *CVPR*.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4060–4073.

Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 2016. Human attention in visual question answering: Do humans and deep networks look at the same regions? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 932–937.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 457–468.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680.

Yash Goyal, Tejas Khot, 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gabriel Grand and Yonatan Belinkov. 2019. Adversarial regularization for visual question answering: Strengths, shortcomings, and side effects. *NAACL HLT 2019*, page 1.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2019. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*.

Allan Jabri, Armand Joulin, and Laurens Van Der Maaten. 2016. Revisiting visual question answering baselines. In *European conference on computer vision*, pages 727–739. Springer.

Kushal Kafle and Christopher Kanan. 2017. An analysis of visual question answering algorithms. In *The IEEE International Conference on Computer Vision (ICCV)*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Leland Gerson Neuberg. 2003. Causality: models, reasoning, and inference, by judea pearl, cambridge university press, 2000. *Econometric Theory*, 19(4):675–685.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Jingjing Pan, Yash Goyal, and Stefan Lee. 2019. Question-conditioned counterfactual image generation for vqa. *arXiv preprint arXiv:1911.06352*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. 2018. Overcoming language priors in visual question answering with adversarial regularization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 1541–1551. Curran Associates, Inc.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 618–626.

Ramprasaath R. Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. 2019. Taking a hint: Leveraging explanations to make vision and language models more grounded. In *The IEEE International Conference on Computer Vision (ICCV)*.

Robik Shrestha, Kushal Kafle, and Christopher Kanan. 2019. Answer them all! toward universal visual question answering models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10472–10481.

Damien Teney, Ehsan Abbasnedjad, and Anton van den Hengel. 2020a. Learning what makes a difference from counterfactual examples and gradient supervision. *arXiv preprint arXiv:2004.09034*.

Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel. 2020b. Unshuffling data for improved generalization. *arXiv preprint arXiv:2002.11894*.

Damien Teney, Peter Anderson, Xiaodong He, and Anton van den Hengel. 2018. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jialin Wu and Raymond Mooney. 2019. Self-critical reasoning for robust visual question answering. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8604–8614. Curran Associates, Inc.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29.

A Appendices

A.1 Implementation Details

The UpDn model uses pretrained Faster R-CNN (Ren et al., 2015) to extract top K object feature embeddings. We set $K = 36$ in our implementation, and the dimension of each object features is 2048. For question embeddings, we preprocess the questions to a maximum of 14 words. The word embeddings are initialized with pretrained GloVe (Pennington et al., 2014) vectors with dimension of 300. A single-layer GRU (Cho et al., 2014) is used to obtain question embedding vectors with the dimension of 512. The dimension of the joint embedding is 2048. The initial learning rate of Adamax optimizer and learning rate decay schedule are followed to the public reimplementation[2]. The entire system is trained end-to-end with both L_{vqa} and L_c. The parameters are initialized from scratch and the random seed is set to 0. The loss weight λ_{vqa} and λ_c are respectively set to 1 and 2. We set batch size to 512. The model developed on the official public Pytorch codebase[3] takes about 5 hours ($\sim$30 epochs) to train on a Nvidia RTX 2080Ti. Both Q-CSS and V-CSS are used to generate (Q, Q^+, Q^-) and (I, I^+, I^-).

A.2 Performance on VQA v2

Model	Expl.	VQA v2 *val*			
		Overall	Y/N	Number	Other
SAN (Yang et al., 2016)		52.41	70.06	39.28	47.84
GVQA (Agrawal et al., 2018)		48.24	31.17	**72.03**	34.65
UpDn (Anderson et al., 2018)		**63.48**	**81.18**	42.14	**55.66**
+AReg (Ramakrishnan et al., 2018)		62.75	79.84	42.35	55.16
+GRL (Grand and Belinkov, 2019)		51.92	-	-	-
+RUBi (Cadene et al., 2019b)		-	-	-	-
+LMH (Clark et al., 2019)		56.35	65.06	37.63	54.69
+LMH+CSS* (Chen et al., 2020)		55.50	61.84	39.82	54.85
+LMH+CSS+GS* (Teney et al., 2020a)		45.11	36.17	38.47	53.70
+LMH+CSS+CL(ours)		57.29	67.27	38.40	54.71
+HINT (Selvaraju et al., 2019)	HAT	62.35	80.49	41.75	54.01
+SCR (Wu and Mooney, 2019)	HAT	62.20	78.90	41.40	54.30

Table 5: Performance comparison on VQA v2 validation split. *indicates the results of our reimplementation.

The results on the VQA v2 are also reported in Table 5 for completeness. We observe that our

[2]https://github.com/hengyuan-hu/bottom-up-attention-vqa

[3]https://github.com/yanxinzju/CSS-VQA

method improves the performance of LMH+CSS from 55.50% to 57.27%. The Gradient Supervision (GS), on the other hand, results in a sharp drop in the performance by 10.39% for LMH+CSS. The phenomenon shows that our approach is more compatible with the i.i.d. setting.