A Self-supervised Strategy for the Robustness of VQA Models

Jingyu Su, Chuanhao Li, Chenchen Jing, Yuwei Wu

A Self-supervised Strategy for the Robustness of VQA Models

Jingyu Su, Chuanhao Li, Chenchen Jing, and Yuwei Wu

Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology (BIT), Beijing, China.
{3120195500, lichuanhao, chenchen.jing, wuyuwei}@bit.edu.cn

Abstract. In visual question answering (VQA), most existing models suffer from language biases which make models not robust. Recently, many approaches have been proposed to alleviate language biases by generating samples for the VQA task. These methods require the model to distinguish original samples from synthetic samples, to ensure that the model fully understands two modalities of both visual and linguistic information rather than just predicts answers based on language biases. However, these models are still not sensitive enough to changes of key information in questions. To make full use of the key information in questions, we design a self-supervised strategy to make the nouns of questions be focused for enhancing the robustness of VQA models. Its auxiliary training process, predicting answers for synthetic samples generated by masking the last noun in questions, alleviates the negative influence of language biases. Experiments conducted on VQA-CP v2 and VQA v2 datasets show that our method achieves better results than other VQA models.

Keywords: Visual Question Answering, Language Bias, Self-supervised Learning.

1 Introduction

The visual question answering (VQA) task requires a model to make comprehensive use of both visual information in images and linguistic information in questions to provide correct answers (Antol et al., 2015). It has attracted lots of interest in the computer vision and natural language processing communities.

During training, most early models learn spurious correlations between question types and answers (i.e., language biases). However, the other parts of the question and the visual information are overlooked, although the model needs to combine the nouns outside of the question type and the image to make inferential region localization. For example, we know a question with the type "What color is the" should be answered by color. It seems that this correlation will provide a set of candidate answers and help the model predict the correct answer. However, if most samples with a question type "What color is the" in the training set have the answer "Red" and the model learned this spurious statistical correlation in training, the model may be able to answer correctly based on "What color is the" alone when we ask "What color is the train?" as depicted in Figure 1a. In this case, the models will ignore the visual information (**Fig. 1(c)**) and the rest parts of the question (**Fig. 1(b)**). These models have been validated to perform poorly on the VQA-CP dataset (Aishwarya et al., 2018) in which the training set has different answer distributions to the test set.

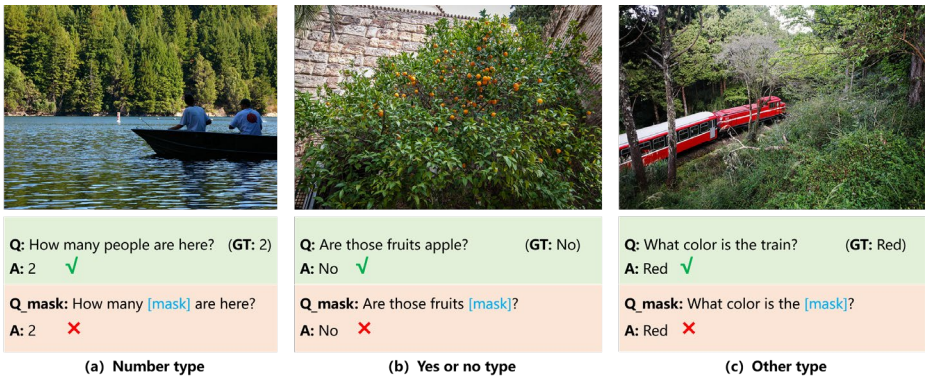

Fig. 1. A model learned language biases cannot respond to the changes (i.e., [mask]) in question-image pairs although it can give the correct answer when the questions are complete. However, when keywords are masked, a robust model will not give the original answers for the lack of key information pointing to a particular answer.

Recent approaches have attempted to address this issue. A popular method is to generate a different input question-image pair and supervise whether the model has learned the bias by seeing if it still predicts the same answer (Zhu et al., 2020; Chen et al., 2020). For example, Zhu et al. (2020) proposed a self-supervised framework, called SSL, that generates samples by randomly replacing the image in original samples and achieves a good result because their framework helps the model

recognize the importance of the visual part in question answering. However, the important regions in questions are not located in the SSL framework. The model still lose sight of some keywords in the question and performs the task well only with the information from the question type in the training process. In this paper, we propose a self-supervised framework based on SSL with an auxiliary task helping the model pay attention to the important parts of the question other than the question type when predicting the answer. In particular, the important parts refer to those keywords which are directly related to visual objects. And we empirically found that they are always the nouns that come later in the sentence.

We introduce an auxiliary self-supervised training process to the conventional training process. The self-supervised training process requires the model to predict the "false" answers for the "wrong" image-question pairs. The "wrong" pairs are generated by masking the nouns and replacing the image, which will be detailed described in **Section 3.2.** Such a self-supervised task can help the model be aware of not only the importance of image but also the differences between partially masked questions and original questions. This encourages the model to pay more attention to the nouns at the end of sentences and consider the information in question more fully when predicting answers.

To sum up, we propose a self-supervised framework based on UpDn (Anderson et al., 2018) to avoid the model learning language biases and design a strategy to generate samples for the self-supervised training progress. Experiments conducted on VQA-CP v2 (Aishwarya et al., 2018) and VQA v2 (Goyal et al., 2017) datasets demonstrate the effectiveness of our method.

2 Related work

Language bias has become a major challenge for VQA researchers. Existing algorithms to solve this problem can be divided into two categories. One is based on highlighting the important visual regions under the guidance of external visual supervision (Ramprasaath et al., 2019; Wu et al., 2019). They are classified as annotation-based methods. These methods work directly but rely heavily on manual labeling.

Another is the no-annotation-based method (Ramakrishnan et al., 2018; Cadene et al., 2019; Remi et al., 2019; Clark et al., 2019; Jing et al., 2020; Zhu et al., 2020; Chen et al., 2020; Abbasnejad et al., 2020). These methods are mainly based on the Up-Down model (Anderson et al., 2018). The Up-Down model performs well on non-inversely distributed datasets like VQA v1 (Antol et al., 2015) and VQA v2 (Goyal et al., 2017). As with other early models like SAN (Yang et al., 2016), MCB (Fukui et al., 2016), and GVQA (Agrawalet al., 2018), its accuracy is decreased on the inverse distributed datasets, VQA-CP (Aishwarya et al., 2018), because of language biases.

A popular solution proposed so far is to generate some auxiliary samples for the training process. For example, SSL (Zhu et al., 2020) generates auxiliary samples by randomly replacing images in the original sample and requires models to distinguish auxiliary samples and origin samples. CSS (Chen et al., 2020) generates auxiliary

samples by masking critical objects in images or words in questions and assigning with different ground-truth answers. These works help the model to be aware of the importance of both images and questions but did not directly indicate which part of the questions should be the focus. The model can still answer only based on the "type" information in questions. In this paper, we generate a group of auxiliary samples by masking the nouns in questions to help models be aware of their importance of them.

Besides generating auxiliary samples, there are also some other effective methods (Clark et al., 2019; Jing et al., 2020). These methods usually use auxiliary branches or delicate structures to eliminate biases.

3 Method

3.1 The Basic VQA Training

A VQA dataset with N samples can be denoted as $D = \{I^i, W^i, A^i\}_{i=1}^N$, where I^i and W^i represent the image and question of the i^{th} sample, and A^i is the corresponding answer. The training target for models is to predict A^i for (I^i, W^i). We build our model upon UpDn (the part of the dashed box in **Fig. 2.**). It is used for basic training. The model uses an image encoder and question encoder, respectively, to embed (I^i, W^i) to (V^i, Q^i) by

$$\begin{cases} V^i = \{o_1^i, ..., o_{n_{J^i}}^i\}, \\ Q^i = \{w_1^i, ..., w_{n_{W^i}}^i\} \end{cases}, \tag{1}$$

where o_j^i is the j-th object feature of I^i, and w_j^i is the j-th word feature of W^i.

The target of basic VQA training is to learn a model which can output the correct answer when inputting the image-question pair. The model can be described as

$$y^i = softmax(\mathcal{F}(V^i, Q^i)), \tag{2}$$

where y^i is a vector with dimensions equal to the total number of answers and represents the probability distribution of the answer prediction. A commonly used loss function for this training is the multi-label soft loss

$$\mathcal{L}_{vqa_{ml}} = -\frac{1}{N}\sum_i^N \left[t_i \log\left(\sigma\left(\mathcal{F}(A^i \mid V^i, Q^i)\right)\right) + (1 - t_i) \log\left(1 - \sigma\left(\mathcal{F}(A^i \mid V^i, Q^i)\right)\right)\right], \tag{3}$$

where $\sigma(\cdot)$ represents the sigmoid function and $\mathcal{F}(A^i \mid V^i, Q^i)$ is the value of answer distribution function $F(A^i) = \mathcal{F}(V^i, Q^i)$ at A^i.

3.2 Sample Generation

We generate two kinds of self-supervised samples for the self-supervised task during the training process. They are named as "question partially masked sample", $\left(V^i, Q^i_{mask}\right)$, and "image randomly replaced sample", $\left(V^i_{rand}, Q^i\right)$. The input of generating process is the original sample (V^i, Q^i). To distinguish it from the synthetic sample, we note it as (V^{i_0}, Q^{i_0}).

To generate $\left(V^i, Q^i_{mask}\right)$, we first tagged the Part-Of-Speech tagging (POS_tag) of each word in Q^{i_0} by NLTK (Bird S et al., 2009). Then we masked the last k nouns in each question to get Q^i_{mask}. And V^i is equal to original V^{i_0}. In the actual experimental environment, we found that multiple masks would lead to unstable training results, since it would make the whole sentence meaningless. It makes no sense to force the model to learn from such an example. To maximize the information available in the sentence, we set $k=1$, which means that only the last noun in the sentence of Q^{i_0} would be masked to get Q^i_{mask}.

The sample generation of $\left(V^i_{rand}, Q^i\right)$ is the same as SSL. The V^i_{rand} means feature of the image which is randomly chosen from the image set to replace the original one (i.e. $V^i_{rand} \in \{V^i\}_{i=1}^N$ and $V^i_{rand} \neq V^{i_0}$). And Q^i is equal to origin Q^{i_0}.

3.3 Self-supervised Training

After generating new samples, we can use them for our self-supervised training in conjunction with basic VQA training. The self-supervised training uses the two kinds of synthetic samples as input and shares the same model used in basic training to predict answers. Different from basic training, the target for this training is not to let the model output the correct answer. While the input is the synthetic samples, the model should no longer output the original correct answer. Therefore, the loss function for self-supervised training should be positively related to the value of the output probability on the origin label. For those two kinds of synthetic samples, the loss function can be defined as

$$\mathcal{L}_q = \frac{1}{N}\sum_i^N P\left(A^i \middle| V^i, Q^i_{mask}\right), \quad \mathcal{L}_v = \frac{1}{N}\sum_i^N P\left(A^i \middle| V^i_{rand}, Q^i\right), \tag{4}$$

where $P\left(A^i \middle| V^i, Q^i\right) = \mathrm{softmax}(\mathcal{F}(A^i \mid V^i, Q^i))$ denotes the predicted probability at the answer annotation. These loss functions mean that the model shall not predict the "correct" answers through "wrong" inputs.

The self-supervised training process can be synchronized with basic training. We only need to let the model predict the label and calculate the loss for the original sample and two synthetic samples respectively, and sum the three losses with different weights. The total loss of training becomes

$$\mathcal{L} = \mathcal{L}_{vqa} + \alpha \mathcal{L}_{v} + \beta \mathcal{L}_{q}, \tag{5}$$

where α and β are hyper-parameters that adjust the weights of self-supervised losses. The $\mathcal{L}_{vqa}$ here can be any VQA loss included mentioned in **Section 3.1**.

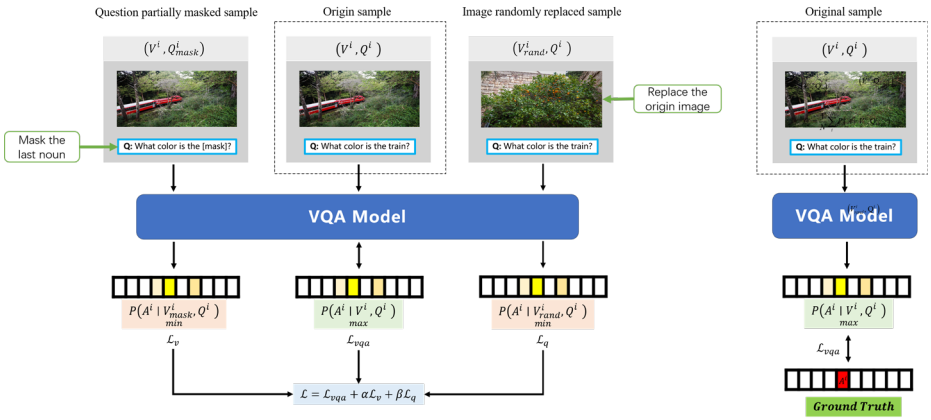

Fig. 2. Self-supervised framework (left) and base model (right). The base model predicts answers for original question-image pairs. And its loss function aims to maximize the probability of producing the answer "Red". In the dashed box we use the original input the same as the base model. The auxiliary branches outside the dashed box answer the question for the auxiliary sample question partially masked input and the image randomly replaced input generate by the progress depicted in **Section 3.2**, respectively. In contrast to the original branch, their loss functions (Eq. (4) in **Section 3.3**) aim to minimize the probability of producing the answer "Red".

Algorithm 1. Model Training with auxiliary self-supervised task

Input: Training sample $D = \{I^i, W^i, A^i\}_{i=1}^{N}$, model $\mathcal{F}$ in Eq. (2), and maximum iteration T.

Output: Updated model $\mathcal{F}$.

1. **while** $t < T$ **do**
2. Encode training sample to the form in Eq. (1).
3. Generate auxiliary samples in **Section 3.2**.
4. Predict answer for original sample and synthetic sample respectively
5. through the same model $\mathcal{F}$.
6. Compute loss for original sample in Eq. (3).
7. Compute loss for synthetic sample in Eq. (4).
8. Backpropagation.
9. **end while**

4 Experiments

4.1 Performance on VQA dataset under Changing Prior

We compared our model with state-of-the-art models on the VQA-CP v2 dataset to evaluate whether our method can effectively avoid language biases problem.

For the first 21 epochs, our model is trained with only $\mathcal{L}_{vqa_{ml}}$, which is set as the multi-label VQA loss, to get a basic ability for the VQA task. Then we add $\mathcal{L}_v$ and $\mathcal{L}_q$ to adjust the model for the last 19 epochs. The hyper-parameter α and β are set to 12.6 and 0.09, respectively. Usually, in our environment, accuracy on the validation set gets the most significant increase at the certain epoch we add the two self-supervised losses.

The results are shown in **Table 1**. The **best** and the second performance are highlighted in each column. **UpDn** + $\mathcal{L}_{vqa_{ml}}$ is different from **UpDn** for its using $\mathcal{L}_{vqa_{ml}}$ replace of the origin VQA loss that **UpDn** used. **UpDn + SSL** and **UpDn + Ours** also use the $\mathcal{L}_{vqa_{ml}}$. It can be observed that our method gets the best overall accuracy and get the best score for the "Yes or No" and "Other" type. For the "Number" type, we achieved a +35.97% improvement over **UpDn** and a +18.03% improvement over **SSL**. It is comparable with **UpDn + LMH + CSS**, which performed best in this particular category.

Table 1. Accuracies (%) of different models on the VQA-CP v2 dataset. The best and second results are bold and underlined respectively.

Method	*Yes or No*	*Number*	*Other*	*Overall*
UpDn	42.27	11.93	46.05	39.74
UpDn + $\mathcal{L}_{vqa_{ml}}$	45.66	16.11	<u>52.27</u>	41.27
UpDn + LMH	72.95	31.90	47.79	52.73
UpDn + CSS	43.96	12.78	47.48	41.16
LMH + CSS	84.37	**49.42**	48.21	<u>58.95</u>
UpDn + SSL	<u>86.53</u>	29.87	50.03	57.59
Ours	**88.62**	<u>47.90</u>	**54.21**	**61.09**

4.2 Performance on traditional VQA dataset

On the VQA dataset without inverse distribution, our model can also reach state-of-the-art performance. It was evaluated on the VQA v2 dataset. The results are shown in **Table 2**. The hyper-parameter α and β are set to 0.1 and 0.5, respectively. After

training with only $\mathcal{L}_{vqa_{ml}}$ for 20 epochs, the model reaches the highest overall accuracy 65.51% for the base. The model at this stage is the same as UpDn + $\mathcal{L}_{vqa_{ml}}$. Then, the accuracy drops a little bit by 0.22% in the last 20 epochs since we add $\mathcal{L}_v$ and $\mathcal{L}_q$ to the total loss. This is an acceptable price to pay for improving the robustness of the model by getting rid of language biases.

Table 2. Accuracies (%) of different models on VQA v2 datasets. The best and second results are bold and underlined respectively.

Method	*Yes or No*	*Number*	*Other*	*Overall*
UpDn	63.79	<u>42.51</u>	55.78	63.79
UpDn + $\mathcal{L}_{vqa_{ml}}$	<u>78.48</u>	**44.98**	<u>57.12</u>	**65.51**
UpDn + LMH	65.06	37.63	54.69	56.35
UpDn + CSS	72.97	40.00	55.13	59.21
LMH + CSS	73.25	39.77	55.11	59.91
UpDn + SSL	-	-	-	63.73
<u>Ours</u>	**80.51**	42.39	**57.20**	<u>65.29</u>

5 Conclusions

In this paper, we have presented a self-supervised strategy to help the VQA model focus on the important nouns in the questions hence avoiding the model learning language biases. In addition, we have proposed an efficient method to generate samples for the self-supervised training. And a series of experiments have verified the effectiveness of the proposed self-supervised method. The operation of masking the last noun, while effective, is simplistic and subjective. A more flexible rule of masking should further improve the accuracy of the model at the expense of more complex structures and more hyper-parameters. This should be considered in future works.

References

1. Antol S., Agrawal A., Lu J., Mitchell M., Batra D., Zitnick C. L., Parikh D. “Vqa: Visual question answering”, in Proceedings of the IEEE international conference on computer vision p. 2425-2433, 2015.
2. Ramakrishnan S., Agrawal A., Lee S., “Overcoming language priors in visual question answering with adversarial regularization”, in Advances in Neural Information Processing Systems, p. 1541–1551, 2018.
3. R. Cadene, H. Ben-Younes, M. Cord, and N. Thome, “Murel: Multimodal relational reasoning for visual question answering”, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, p. 1989-1998, 2019.

4. X. Zhu, Z. Mao, C. Liu, P. Zhang, B. Wang, and Y. Zhang, “Overcoming language priors with self-supervised learning for visual question answering”, International Joint Conference on Artificial Intelligence, p. 1083-1089, 2020.
5. Chen L., Yan X., Xiao J., Zhang H., Pu S., Zhuang Y., “Counterfactual samples synthesizing for robust visual question answering”, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, p. 10800–10809, 2020.
6. Clark C., Yatskar M., Zettlemoyer L., “Don’t take the easy way out: Ensemble based methods for avoiding known dataset biases”, Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, p. 4060–4073, 2019.
7. Abbasnejad E., Teney D., Parvaneh A., Shi J., Hengel A. V. D., “Counterfactual vision and language learning”, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, p. 10044-10054, 2020.
8. Jing C., Wu Y., Zhang X., Jia Y., Wu Q., “Overcoming language priors in vqa via decomposed linguistic representations”, AAAI Conference on Artificial Intelligence Vol. 34, No. 07, p. 11181-11188, 2020.
9. Anderson P., He X., Buehler C., Teney D., Johnson M., Gould S., Zhang L., “Bottom-up and top-down attention for image captioning and visual question answering”, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, p. 6077–6086, 2018.
10. Yang Z., He X., Gao J., Deng L., Smola A., “Stacked attention networks for image question answering”, in Proceedings of the IEEE conference on computer vision and pattern recognition, p. 21-29, 2016.
11. Fukui A., Park D. H., Yang D., Rohrbach A., Darrell T., Rohrbach M., “Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding”, Empirical Methods in Natural Language Processing, p. 457–468, 2016.
12. Agrawal A., Batra D., Parikh D., Kembhavi A., “Don’t just assume; look and answer: Overcoming priors for visual question answering”, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4971-4980, 2018.
13. Selvaraju R. R., Lee S., Shen Y., Jin H., Ghosh S., Heck L., Batra D., Parikh D., “Taking a hint: Leveraging explanations to make vision and language models more grounded”, in Proceedings of the IEEE International Conference on Computer Vision, p. 2591–2600, 2019.
14. Wu, J., & Mooney, R. J., “Self-critical reasoning for robust visual question answering”, In Advances in Neural Information Processing Systems, p. 8601–8611, 2019.
15. Cadene, R., Dancette C., Cord M., Parikh D., “Rubi: Reducing unimodal biases for visual question answering”, Advances in neural information processing systems, p. 841–852, 2019.
16. Goyal Y., Khot T., Summers-Stay D., Batra D., Parikh D., “Making the v in vqa matter: Elevating the role of image understanding in visual question answering”, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, p. 6904–6913, 2017.
17. Bird S, Klein E, Loper E., Natural language processing with Python: analyzing text with the natural language toolkit. “O’Reilly Media, Inc.”, 2009.