A Critical Robustness Evaluation for Referring Expression Comprehension Methods

Zhipeng Zhang*
zhipeng.zhang@mail.nwpu.edu.cn
Zhimin Wei*
ZhiminWei@mail.nwpu.edu.cn
Peng Wang
peng.wang@nwpu.edu.cn

The National Engineering Laboratory for Integrated
Aero-Space-Ground-Ocean Big Data Application Technology,
School of Computer Science,
Northwestern Polytechnical University

Abstract

Referring Expression Comprehension (REC) is a crucial task in visual reasoning that requires models to accurately identify target objects indicated by natural language expressions. Researchers have focused on the performance of models on the COCO dataset (RefCOCO/RefCOCO+/RefCOCOg) and employed various training strategies to improve performance scores. However, there is a lack of robustness evaluation and analysis among these works due to the absence of an evaluation metric and dataset benchmarks for comparison. In this work, we propose a novel dataset and benchmark for the word-level adversarial robustness of Referring Expression Comprehension task. We also evaluate the robustness experiments on several previous strong methods. The coda and dataset will be available at https://github.com/tujun233/C2C-R

1 Introduction

Referring Expression Comprehension (REC) [9, 29] is a challenging task that aims to accurately identify target objects within images based on natural language expressions. This task requires models to understand the relationship between visual and linguistic information and to generate accurate and precise predictions. To achieve this goal, the models need to effectively comprehend both the expression and image.

Initially, detection approaches [5, 8, 13, 14, 17, 30] were commonly used for inference in REC models. They relied on a CNN-based detector [7, 18] to produce numerous candidate proposals that potentially contained the target object. The models then utilized natural language text to filter out the most appropriate candidate region. However, because of inherent limitations in image detectors, one-stage/encoder-decoder approaches [10, 12, 22, 25, 26] have been increasingly popular in recent years. These methods directly fuse natural language expressions and image features to generate the final candidate regions, bypassing the traditional two-stage approach of generating region proposals. Additionally, various strategies have been implemented [22, 24, 26, 28, 31], such as multi-step reasoning, reinforcement

* means the same contributions for these authors.

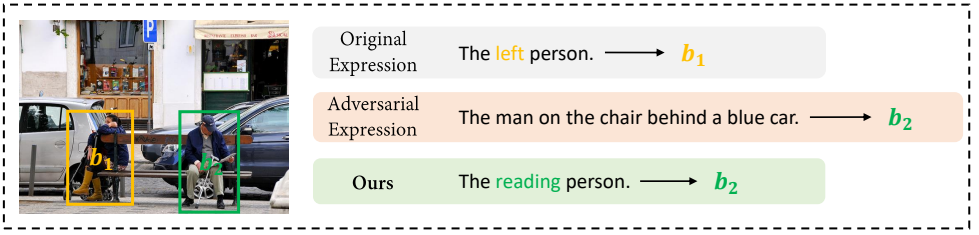


Figure 1: The comparison between existing robustness evaluation and ours. Previous evaluation methods rewrite a new expression. In this work, we show our word-level robustness evaluation of adversarial words. The b_i is the different objects.

learning, and attention, to improve comprehension ability during the fusion and reasoning process. Meanwhile, vision language pre-training models [22, 28] have also become popular, leading to significant performance improvements across several datasets. These models are trained or fine-tuned on multiple datasets to achieve better fit. The current state-of-the-art model for REC tasks, OFA [22], leverages large-scale multi-task vision-language alignment pre-training to establish an encoder-decoder baseline model. Subsequently, the model is fine-tuned on individual REC task datasets independently to improve accuracy, which incurs significant costs and introduces uncertainty.

However, while researchers are focused on achieving optimal performance, there are some potential limitations to consider. The primary metric used in the current REC review entails calculating the Intersection over Union (IoU) score for the predicted and ground-truth bounding boxes. This evaluation approach is inherited from the field of computer vision detection and recognition and primarily reflects the model's detection capability rather than its ability to comprehend both images and text accurately. Some studies [1, 2] have suggested that previous models could be exploiting strong biases in these datasets and tried to introduce the robustness study to improve. Akula [1] was aware of the impact of text robustness on the results and incorporated revised adversarial text only in the RefCOCOg portion of the test to validate and enhance their model's performance. However, as demonstrated in Figure 1, they tend to rely on color words to locate target regions, leading to potential inaccuracies in their predictions. The revised expression does not provide a targeted means to evaluate the model's ability to handle word variations or the potential shortcomings of the language model. This undermines the true purpose of the task and raises doubts about the actual progress made. As a result, the robustness of the model is an essential parameter, particularly when used for robotic navigation, detection, and localization in real-life settings where language is diverse and complex. In such scenarios, the model's ability to comprehend natural language expressions accurately can have a significant impact on the overall performance of the system. Thus, the development of accurate and robust REC models is critical for advancing the field of visual reasoning. Undertaking robustness benchmark tests on AI models can aid in identifying and addressing potential issues, thereby improving model dependability. However, in the REC task, there is a lack of corresponding complete datasets and benchmarks to evaluate model robustness.

In our work, we re-labeled the common datasets, Refcoco/Refcoco+/Refcocog. We focus on word level and take text fluctuations in natural language descriptions to test the robustness of the model. Of course, we only re-annotated the data samples that the original model was able to predict correctly. Specifically, we created two sub-robustness datasets: Change-

Change (C2C) and Change-Remain (C2R). In C2C, changing one or more keywords in the text description leads to a change in the ground truth anchor box position; whereas in C2R, changing one or more words in the text description does not alter the true target box position. To evaluate the performance of the tested models, we used 0.5 IoU ratio to compare the predicted box and ground-truth box. A predicted box with an IoU ratio greater than 0.5 is considered accurate, and the correct accuracy is then calculated. By analyzing these metrics, we can determine the effectiveness of the tested models in accurately predicting the target region of the referred object under different conditions.

In conclusion, our study makes the following main contributions:

(1). We proposed an examination of vision-language (VL) model robustness and meticulously developed both a dataset and benchmark for evaluating word-level robustness in referring expression comprehension tasks.

(2). We conducted a thorough evaluation and analysis of various representative mainstream latest methods for robustness. Our research has the potential to advance the works on referring expression comprehension, providing a foundation for further exploration in this field.

2 Related Work

2.1 Referring Expression Comprehension

Referring Expression Comprehension (REC) is a visual-language cross-modal understanding problem. It aims to detect the target object described by a natural language expression in an image. Most previous methods generate several region proposals by yolo or Faster RCNN in the first stage. The second stage is to retrieve the objected region matched with the input expression by calculating the similarity. Kazemzadeh [9] was the first to propose the refcoco,refcoco+ dataset building on the coco dataset. Mao [14] then presented the refcocog data annotation collection. These three datasets are all annotated on the basis of coco 2014 images. However, refcocog has longer annotated text and complex relationships between multiple targets. Yang [25] puts forward a one-stage model that fuses an expression's embedding into a YOLOv3 [16] object detector augmented by spatial features, and then it uses the merged features to localize the corresponding region. The model OFA unifies modality, task, and structure. It unifies multimodal understanding and generation tasks into a simple Seq2Seq generative framework. OFA covers downstream tasks across multiple scenarios such as multimodal generation, multimodal understanding, image classification, natural language understanding, and text generation. Among them, SOTA results are obtained on the coco dataset. However, existing evaluation metrics, IoU anchor calculation, only tend to express the accuracy performance of these models on that dataset and ignore robust safety analysis.

2.2 Robustness Evaluation

The robustness [11, 21] of the model mainly refers to the stability and effectiveness of the system despite certain noise fluctuations or slight deviations from the control quantity. Generalization ability means that the network model obtained from a finite sample has a good predictive ability for other variable domains as well. In computer vision, robustness evaluation [6] involves adding noise to analyze a model's stability or conducting adversarial attacks

by training adversarial samples to enhance model security. A model with good robustness can perform better on new or noisy data. Narodytska [15] presents the robustness verification problem for Binarized Neural Networks (BNNs), the first exact Boolean representation of deep neural networks, independent of the network structure. Xiang [23] studied the problem of robust verification of Multi-Layer Perception (MLP) machines by estimating the set of outputs of MLP from a large number of simulation results and performing robustness verification. Wang [20] introduced poorly specified Gaussian noise injected into ResNet to test the average experimental results to verify the robustness of the model. In conjunction with the study of robustness problems such as adversarial attacks, we perform robustness tests for possible noise attacks (text-guided noise) in the field of multimodal referring expression comprehension, such as word variations. Duboue [4] defined a methodology to evaluate the generation of referring expressions algorithm. Akula2020words [1] recognized the issue of robustness in referring expression comprehension (REC) and addressed it by adding adversarial text annotations to a subset of data in the RefCOCOg test set. However, these approaches did not include a complete benchmark or comprehensive dataset that incorporates several common datasets, methods, and image coordinate changes.

3 Dataset

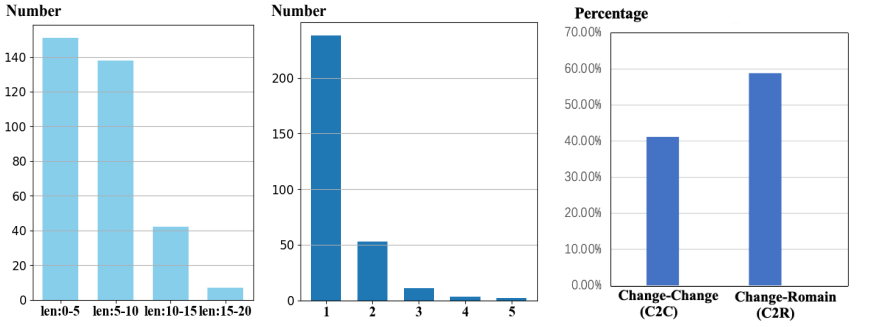
Refcoco [29], Refcoco+ [29] and Refcocog [14] are three referring expression datasets with images and reference objects selected from coco images. The language is selected from coco object detection annotations into 80 object classes. Refcoco+ is similar to Refcoco but prohibits the use of absolute positional words, so it clearly requires more effort. Specifically, Refcoco+ expressions do not contain words with positional relation properties. For example, "on the left" describes the position of the object in the image. Queries in Refcocog are generally longer than those in Refcoco and Refcoco+: the average lengths of Refcoco, Refcoco+, and Refcocog have an average length of 3.61, 3.53, and 8.43, respectively. It is worth mentioning that there are a lot of simple reference expressions in refcoco and refcoco+, such as "man", "the woman", "guy", etc. And refcoco+ simply removes the absolute position words, like "on the far left".

In the data labeling phase, we screened Refcoco, Refcoco+, and Refcocog. We first filter the original images, which include simple images, single words or phrase text descriptions. It is difficult for us to produce robust changes in the model in the face of these simpler text descriptions and images. As shown in table 1 and 2, we select multiple target images and complex text descriptions, which include text descriptions of relationships to salient regions in the image, feature descriptions such as color location, semantic information mining, etc.

To ensure the accuracy of our data, we implemented a calibration process in which each group consisted of two professional visual grounding researchers. These researchers worked together to label and check each other's work, thereby reducing errors and ensuring consistency. Only when his re-annotations enable others to find the correct answers are successful cases. This calibration process helped to validate the accuracy of the data used in our experiments, which ultimately improves the overall quality and reliability of our results.

In total, we generated 338 images from 1,364 images filtered for annotation. Among them, in refcoco we generated 122 annotated images from 676 referring expression cases, in refcoco+ we generated 103 annotated images from 304 images, and in refcocog we generated 119 annotated cases from 384 raw cases.

As shown in Figure 2(b), we changed the key words causing the image target to change,



(a) The length of captions (b) The number of changed words (c) The percentage of C2C and C2R in robustness test dataset.

Figure 2: The statistical results about our robustness test dataset. We visualize the following in our robustness evaluation dataset: the caption length, the number of words changed per expression, and the ratio of C2C and C2R.

	Change-Change	Change-Remain
Person	49.5%	66.4%
Animal	10.5%	9.4%
Food	9.5%	6.7%
Vehicle	4.5%	2.7%
Furniture	7.0%	2.7%
Outdoor	3.0%	0.7%
Indoor	13.5%	9.4%
Clothes	2.5%	2.0%

Table 1: The different scene category statistics in our robust dataset.

	Change-Change	Change-Remain
Position	28.50%	10.27%
Color	26.57%	15.07%
Wearing	5.31%	9.59%
Object	10.14%	27.40%
Person	7.73%	15.07%
Attribute	11.61%	12.32%
Action	3.86%	8.22%
Noun	6.28%	2.06%

Table 2: The word-level adversarial word attribute statistics in our robust dataset.

which is where we will redraw the location coordinates (x,y,w,h) of the new target box using IoU, where (x,y) are the coordinates of the upper-left corner of the box and (w, h) is the size of the box. Of course, for the reliability of the experiment, we use two-person interactive labeling. On the one hand, we annotate the linguistic noise changes, and on the other hand, we test whether the target frame can be found correctly. The two researchers test each other's annotations, resulting in reliable human-level annotations. The intersection over union (IoU) ratio of the prediction box and ground-truth box is greater than 0.5, and the Accuracy is calculated.

As shown in the example provided in Figure 3, we tested the robustness of the model primarily by making minor changes to one or more words. While the changes made in our experiments may seem relatively small when compared to other assessments of model robustness, such as the ones conducted in prior research [1], they were crucial in evaluating the model's linguistic comprehension and confirming its linguistic biases. Our emphasis was on assessing the linguistic inertia of the word-level adversarial robustness validation model.

4 Metric

Each method was tested on the robustness test dataset, and the results were compared with the target region using Intersection over Union (IoU) calculation. A repeat region with an

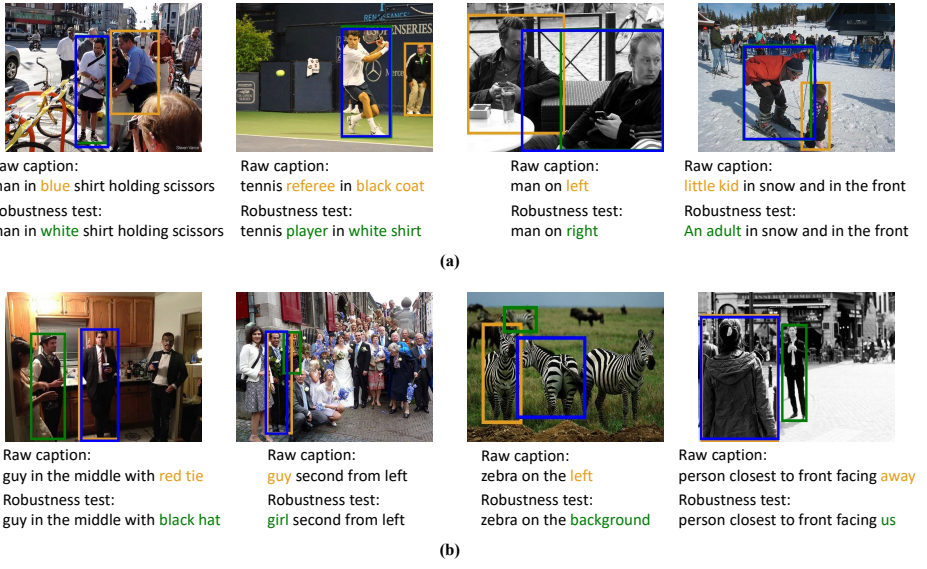


Figure 3: Quantitative analysis of our dataset. We also test the performance of the SOTA method, OFA. The green, yellow, and blue boxes represent the ground truth boxes of the raw text, robustness text, and the results of OFA models.

IoU score greater than 50% was considered correct. We used the more intuitive correctness and error rates as our robustness test metrics:

(1) If the text changes, but the target anchor box position remains the same and the model output does not change, it is considered true.

(2) If the text changes, but the target region remains the same, and the model output changes and does not match the target region IoU calculation, it is considered false.

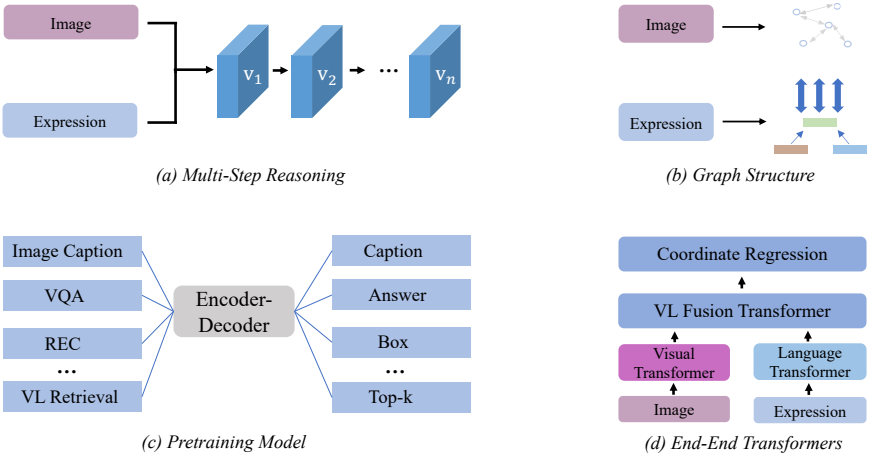
(3) If the text changes, and the target anchor box changes, but the model output produces an IoU value consistent with the new ground truth target anchor box, it is considered true.

(4) If the text changes, and the target region changes, but the model output does not match the new ground truth target region IoU calculation, it is considered false.

$$Accuracy\_Rate = \frac{r_i}{r_i + f_i}, \tag{1}$$

where r_i and f_i mean the right/failure number of the test method in the dataset.

Consequently, we obtained the Accuracy rates for each of the C2C and C2R datasets, resulting in total performance measures in Table 3. It is worth noting that we would like to emphasize that the final correctness and error rate calculations for each method are based on the sum of the number of correct and error samples for both datasets. Therefore, the final percentage calculations are not equal to the direct addition of the two correctness rates due to the differing proportions of the datasets.



(a) Multi-Step Reasoning *(b) Graph Structure*

(c) Pretraining Model *(d) End-End Transformers*

Figure 4: Schematic diagram of the framework of the four mainstream approaches. The multi-task and multi-data pre-training models represented by OFA have become the state-of-the-art pipeline in the field.

5 Experiment

In this section, we selected the latest representative methods and evaluated them through experiments, including multi-step reasoning, graph structure, pre-training models, and end-end transformers. Figure 4 displays the four mainstream frameworks denoted as (a), (b), (c), and (d), with Resc, SGMN, OFA, and TransVG representing the respective state-of-the-art methods chosen under each framework. We also compared and analyzed their results in section 5.1. To ensure fairness, we evaluated the adversarial robustness of the referring expression comprehension task on the same GPU hardware for all experiments.

5.1 Comparison

Table 3 presents the results of our robustness test dataset using the four framework methods. Resc is a classical multi-step inference method, SGMN utilizes graph networks for referring expression comprehension, TransVG employs an end-to-end transformer structure, while OFA is a pre-training vision language model with multi-task and multi-data composition. Notably, OFA is currently the top-performing method on several referring expression comprehension datasets. ERNIE-ViL incorporates structured knowledge obtained from scene graphs to learn visual language pre-training. RefTR, on the other hand, is a one-stage transformer pre-training framework combining multiple task datasets. To ensure maximum accuracy, we tried to test the model using its official release version. The results indicate that the pre-training model has significantly better robustness performance than the others.

OFA is a framework that unifies various cross-modal and unimodal tasks within a simple pre-training encoder-decoder learning framework, achieving state-of-the-art performance across tasks such as visual localization, image captioning, and VQA. It was trained on 20 million publicly available image-text pairs and follows instruction-based learning in both its pre-training and fine-tuning phases. In contrast, Resc, a one-stage visual-textual alignment method that uses ResNet [7] and BERT [19] as feature backbones, demonstrated poorer per-

	Methods	Types	Change-Change (C2C)	Change-Remain (C2R)	Final Robustness
w/o Pretrain	Resc [26]	Multi-Step Reasoning	38.19%	38.13%	38.17%
	SGMN [24]	Graph Network	40.20%	43.17%	41.42%
	TransVG [8]	Transformer-based	41.72%	53.24%	46.45%
Pretrain	ERNIE-ViL [28]	Knowledge Scene Graph	55.28%	69.06%	60.95%
	RefTR [10]	Rec+Res Pretrain	57.79%	71.22%	63.31%
	OFA [23]	**Multi-task VL Pretrain**	**64.32%**	**82.01%**	**71.60%**

Table 3: We performed an analytical validation of several representative methods on our proposed robustness dataset. We also focus on verifying the critical robustness of currently popular pre-training paradigms. The results are the accuracy rates on both the C2C and C2R datasets and the final overall dataset. The Rec and Res are the referring expression comprehension and referring expression segmentation tasks. The OFA is the current SOTA method.

formance in our experiments due to its fixed multi-step inference structure. Furthermore, the graph network structure and the end-to-end transformer structure also perform worse in terms of robust performance. TransVG is a transformer-based structure that incorporates a visual transformer and text transformer to encode two modal features. The overall transformer fusion layer then fuses these features to predict the coordinates of the target frame. As one of the best-performing transformer-based models, it has demonstrated competitive performance by effective fusion of visual and textual information. SGMN is a notable work that utilizes graph networks for referring expression comprehension. It converts visual and textual inputs into two separate graph networks using scene graph representations. The AttendNode units are then used to infer key target regions from these networks. Overall, the three structures (Resc, TransVG, and SGMN) are not even as good as the 50% random probability of being correct. RefTR, proposed to unify pre-training using Ref two specific tasks, referring expression comprehension REC and referring expression segmentation RES [27, 30], because of the similarity of the two tasks' data. ERNIE-ViL, on the other hand, proposes to add knowledge pre-training enhancement. ERNIE-ViL, RefTR and OFA, which are based on pre-training frameworks, have achieved better results. However, in the face of word-level robustness studies, these VL models still suffer from the potential inertia of language models resulting in insufficient performance.

	Position	Color	Wearing	Object	Person	Attribute	Action	Noun
Change-Change	75.32%	71.63%	71.61%	60.35%	60.99%	74.03%	**34.91%**	43.99%
Change-Remain	78.98%	85.68%	92.50%	84.25%	85.68%	70.80%	**56.24%**	65.08%

Table 4: The different attribute words robustness on the pre-training model OFA. The worst performance is highlighted in **bold**.

6 Discussion

As shown in Figure 5, visualizing the results allows us to gain insights into the effectiveness of these different pipeline methods in terms of their robustness. By observing the ground truth boxes represented in blue and comparing them to the green, yellow, and red boxes representing the model predictions, we can assess the models' ability to achieve accurate predictions in robustness testing. Therefore, visualizations help determine the robust performance of the tested methods and their potential for real-world applications. By combining Table 3 and Figure 4, we infer that the pre-trained models exhibit significantly better robust-

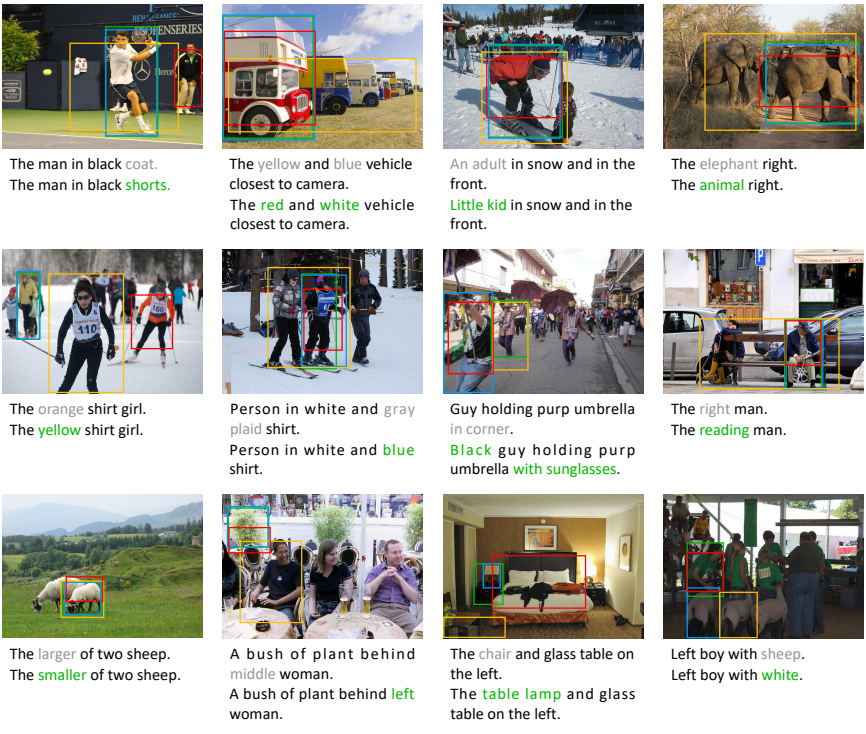


Figure 5: Random examples from our proposed robustness REC dataset. We present the visualization of the results obtained from different methods. The ground truth boxes of the robustness test are represented in blue, while the green, yellow, and red boxes depict the outcomes of the testing models (OFA, Resc, TransVG).

ness than other paradigms in the REC task. This improvement may be due to the noise boost provided by the large amount of pre-trained data. We observed that SGMN outperforms ReSC and is competitive with TransVG, indicating that parsing graph structures improve word-level adversarial robustness. Furthermore, the results from OFA and RefTR highlight the benefits of pre-training with transformer structure on multiple tasks.

We conducted a comprehensive analysis on the category robustness of the state-of-the-art performance model OFA, which is presented in Table 4. Our findings indicate that the performance of OFA is limited by its adversarial robustness in action understanding and counting. This suggests that current image-text-based systems struggle to fully comprehend the movements of characters in an image, which may require additional information from videos to enhance and strengthen image understanding. Furthermore, our analysis shows that OFA's pre-training model finds it challenging to distinguish complex quantities that correspond with multiple targets.

7 Conclusion

In this paper, we present a critical examination of the robustness of referring expression comprehension tasks to fluctuations at the word level. We specifically verify the effectiveness of

recent mainstream method paradigms and investigate whether current vision language pre-training paradigms rely on the tendency of language models to take shortcuts. Our experiments demonstrate that the current model's robustness is still insufficient, indicating a need for further research and analysis.

8 Acknowledgements

This work was supported by National Key R&D Program of China (No. 2020AAA0106900), the National Natural Science Foundation of China (No. U19B2037, No. 61876152), Shaanxi Provincial Key R&D Program (No. 2021KWZ-03), and Natural Science Basic Research Program of Shaanxi (No. 2021JCW-03).

References

[1] Arjun R Akula, Spandana Gella, Yaser Al-Onaizan, Song-Chun Zhu, and Siva Reddy. Words aren't enough, their order matters: On the robustness of grounding visual referring expressions. *arXiv preprint arXiv:2005.01655*, 2020.

[2] Volkan Cirik, Louis-Philippe Morency, and Taylor Berg-Kirkpatrick. Visual referring expression recognition: What do systems actually learn? *arXiv preprint arXiv:1805.11818*, 2018.

[3] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1769–1779, 2021.

[4] Pablo Ariel Duboue, Martin Ariel Domínguez, and Paula Estrella. Evaluating robustness of referring expression generation algorithms. In *2015 Fourteenth Mexican International Conference on Artificial Intelligence (MICAI)*, pages 21–27. IEEE, 2015.

[5] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016.

[6] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[8] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. Modeling relationships in referential expressions with compositional modular networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1115–1124, 2017.

[9] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.

[10] Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. *Advances in neural information processing systems*, 34:19652–19664, 2021.

[11] Xingchen Li, Long Chen, Jian Shao, Shaoning Xiao, Songyang Zhang, and Jun Xiao. Rethinking the evaluation of unbiased scene graph generation. *arXiv preprint arXiv:2208.01909*, 2022.

[12] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 10034–10043, 2020.

[13] Ruotian Luo and Gregory Shakhnarovich. Comprehension-guided referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7102–7111, 2017.

[14] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016.

[15] Nina Narodytska, Shiva Kasiviswanathan, Leonid Ryzhyk, Mooly Sagiv, and Toby Walsh. Verifying properties of binarized deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[16] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[17] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[18] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[20] Bao Wang, Zuoqiang Shi, and Stanley Osher. Resnets ensemble via the feynman-kac formalism to improve natural and robust accuracies. *Advances in Neural Information Processing Systems*, 32, 2019.

[21] Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, et al. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. *arXiv preprint arXiv:2302.12095*, 2023.

[22] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022.

[23] Weiming Xiang, Hoang-Dung Tran, and Taylor T Johnson. Output reachable set estimation and verification for multilayer neural networks. *IEEE transactions on neural networks and learning systems*, 29(11):5777–5783, 2018.

[24] Sibei Yang, Guanbin Li, and Yizhou Yu. Graph-structured referring expression reasoning in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9952–9961, 2020.

[25] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4683–4693, 2019.

[26] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving one-stage visual grounding by recursive sub-query construction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 387–404. Springer, 2020.

[27] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10502–10511, 2019.

[28] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3208–3216, 2021.

[29] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016.

[30] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1307–1315, 2018.

[31] Zhipeng Zhang, Zhimin Wei, Zhongzhen Huang, Rui Niu, and Peng Wang. One for all: One-stage referring expression comprehension with dynamic reasoning. *Neurocomputing*, 518:523–532, 2023.