Robustness of State-of-the-Art Phrase Grounding Models using Corrupted Linguistic Dependency Structures

TECHNISCHE UNIVERSITÄT DARMSTADT

Group D:
Finn Radatz
Lars Reining
Doğukan Bağcı
Hannah Willkomm
Tim-Michael Krieg

Deadline: Monday 8th July, 2024

1. Introduction

In recent years, multimodal input models have gained considerable attention in academic literature due to their broad applicability. Unlike unimodal models that rely on information from a single modality, multimodal models can incorporate data from multiple sources, such as text-image, text-video-audio, and more. This has resulted in significant advancements in various areas, including image classification [22], object detection [9], and text-to-image-generation [23].

A key challenge in machine learning research is the lack of resilience to corruption in existing unimodal models. [7, 15, 33]. Despite the advantages of using multiple modalities, the issue of corruption robustness persists, with limited prior research focusing on out-of-domain or corruption robustness in the multimodal setting, particularly in phrase grounding [1, 26]. This raises concerns about the suitability of such models for safety-critical applications.

Our study evaluates the robustness of state-of-the-art phrase grounding models using corrupted linguistic dependency structures. This form of corruption allows us to assess the influence of structure on model performance. By analyzing models of varying sizes and architectures on our corrupted phrase grounding dataset, we aim to explore the interaction between robustness and accuracy in different models. We hypothesize that the accuracy of grounding models tends to decrease as the textual prompt contains less syntactic structure.

2. Related Work

Adversarial Robustness A closely related field to corruption robustness is the defense of machine learning against adversarial attacks. In an adversarial attack, an adversary introduces small distortions to one or more modalities in order to confuse the models. These attacks are typically so subtle that humans are largely immune to them, but research has shown that models are highly vulnerable to such attacks [32, 35, 25, 18, 3]. One popular defense strategy against adversarial attacks is adversarial training, where the training dataset also includes adversarial examples [16].

Corruption Robustness in Related Fields In contrast to assessing adversarial robustness, which focuses on measuring model accuracy against malicious attacks, corruption robustness evaluates model behavior under corrupted data. Although fields like computer vision and natural language processing have been leveraging machine learning methods for a longer duration, the issue of unreliable model predictions in these domains persists. To address corruption robustness in image classification, the well-established ImageNet-C dataset [7], derived from the original ImageNet-1k dataset [24], contains images corrupted by 15 different types of corruptions, such as various levels of gaussian noise. Additionally, in a related context, [33] introduce modifications to the GLUE dataset [34] by removing certain word groups, like noun phrases. Parallel to our methodology, [29, 31] conduct experiments by shuffling word order and observing performance degradation.

Corruption Robustness of multimodal models [37, 21] examine the resilience of models to text and image alterations. [26] delves into the impact of word shuffling on the Visual Question Answering (VQA) task, highlighting the vulnerability of current models to changes in word order. [1, 4] investigate models' robustness in visual grounding of referring expression models by permuting word order, although this approach differs from the syntactic analysis used in our study.

3. Main Contribution

1. We are the first to provide evidence on the robustness of models against manipulations on different levels of the syntactic structure, allowing a more fine-grained insight into the relevance of structure on SOTA phrase grounding models.

2. We provide an open source annotations variation for the flickr30k entities test dataset, containing corruptions on different levels of the syntactic structure of the textual prompts.

4. Experimental Setup

Task and Dataset In phrase grounding, given a phrase, a model tries to predict which part of the image corresponds to which part of the phrase. It does this by marking each part of the phrase as a bounding box in the image. We evaluate the performance of our model zoo using the Recall@k metric. This metric measures if the top-k predictions of the model contain the correct bounding box. A bounding box is considered correct when the intersection over union between the predicted and any ground-truth box for a specific phrase is over a threshold of 0.5, as used in [9]. To perform our experiments, we use the flickr30k entities dataset [20], which is an extension of the original flickr30k dataset [36]. The extension adds manually annotated bounding boxes that the models can learn.

Corruptions Here, we provide an open-source method for corrupting sentences with systematic control of the degree to which the original sentence structure is preserved. Therefore, our approach allows for the quantitative analysis of the influence of syntactic structure on the performance of different phrase grounding models.

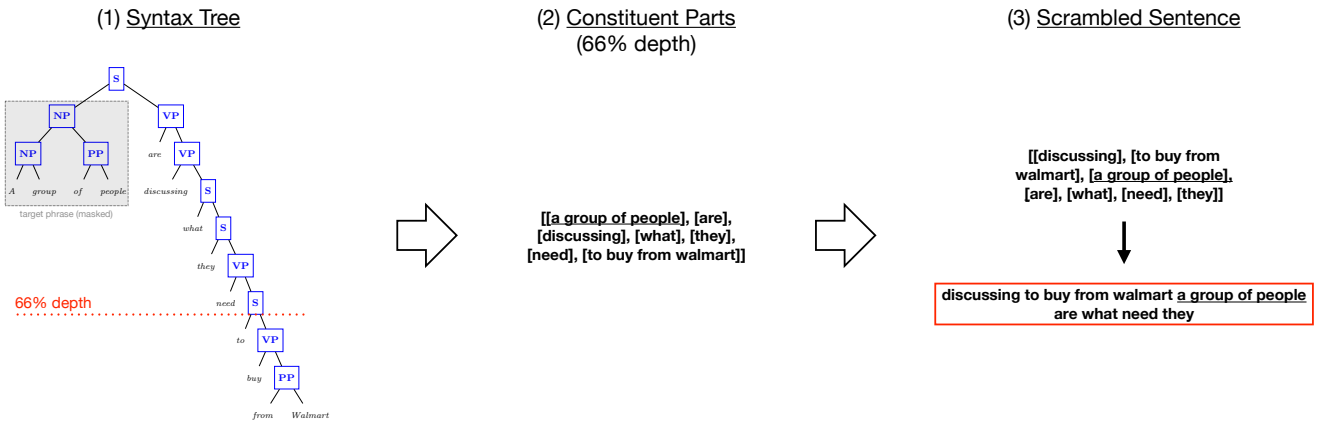

Figure 1: **Manipulation pipeline for sentence scrambling using 66% depth level**. For each sentence, we constructed its syntactic tree (1), additionally masking the target phrases to ensure that they would be preserved in their original order after scrambling. Using the resulting constituent parts (2) at the selected depth (here: 66% depth), the sentence is scrambled (3).

As a first step, we constructed the syntactic tree for each sentence contained in the flickr30k entities dataset [20] using the Berkeley Neural Parser [10, 11]. A syntactic tree is a graphical depiction of the syntactic structure of a sentence that illustrates the hierarchical relationship between different parts of a sentence, i.e., phrases or words, and how they combine to form the sentence. Each node in the syntactic tree represents a phrase or word, whereas tree branches indicate the relationships between the different parts.

We filtered all sentences of the flickr30k entities dataset [20] with a maximum tree height smaller than two to ensure that the effect of the scrambling would be profound for all applied manipulations resulting in the dataset "Filtered." All sentences were converted to lowercase, and punctuation was removed so that there would be no indication of the original sentence structure. Additionally, to ensure that the target phrases would remain unmodified, those phrases were masked for the construction of the syntactic trees.

Using the syntactic trees, we scrambled each sentence on different depths of the syntactic tree. We scrambled on a specific depth d by retrieving the constituent parts at depth $min(d, h_i)$ for each branch of the syntactic tree, where h_i refers to the height of the current tree branch, and then shuffling the original sentence s based on these constituent parts, i.e., sub-sentences s_i of sentence s.

We decided to shuffle on four different depths, which we defined as a percentage relative to the maximum height of the syntactic tree of each sentence, i.e., "Scrambled 33% Depth", "Scrambled 50% Depth", "Scrambled 66% Depth", "Scrambled Word Level" where "Scrambled Word Level" corresponds to "Scrambled 100% Depth". A higher percentage results in less preservation of the original sentence structure, also exhibited by the average Damerau–Levenshtein distances of the different manipulation categories (see Appendix C.1). For a visualization of the manipulation pipeline for a single sentence, refer to Figure 1.

This corruption approach resulted in 18648 scrambled sentences where 5.51% (1027 sentences) remained unmodified.

Model Zoo For our analysis of the effects of corruption, we utilized five models from three different model families. We specifically selected these model families based on their superior accuracy on the flickr30k entities dataset [20], as indicated by the leader board on paperswithcode[1]. The selected model families are GLIP[13], MDETR [9], and FIBER-b [5]. It's important to note that our model zoo only includes models with publicly available weights. Additionally, we specifically considered models that utilized the MDETR annotations to ensure better comparability due to project time constraints. For more detailed information on the various models, refer to Appendix C.2.

5. Results and Analysis

5.1. Analysis

We measured each model and scrambling depth's performance using Recall@1 on the respective dataset. Other recall metrics (e.g., Recall@5) provide similar results (see Appendix C.3).

The relative robustness of a specific model was measured by computing the difference between its performance on the unscrambled dataset and its performance on the most scrambled (scrambled at word-level) dataset. Therefore, negative values indicate that the model performs worse on the scrambled dataset. The more negative the robustness value gets, the less robust the model is.

Finally, we grouped the phrases according to the categories provided in the original dataset and evaluated the models' relative robustness across different categories.

5.2. Results

Before manipulating the data, we checked that we could reproduce the reported results on the original dataset. The performance of our models as measured by their Recall@1 can be seen in the "unfiltered" column of figure 2a. The recall of all models aligns with the reported recalls in [5, 9, 13].

In the next step, we ensured that filtering out low syntactic depth sentences does not affect model performance. This can also be seen in Figure 2a, where the recall values on the filtered and unfiltered datasets are almost identical.

Evaluating the models on scrambled sentences revealed that scrambling of sentences reduces the performance by between 7.2% and 8.2%. We were also able to show that the reduction of performance depends strongly on the depth of the syntactic tree we shuffled on. While scrambling at a higher level (i.e., keeping larger phrases fixed) results generally only in a smaller decrease in performance, scrambling at the word level results in the highest performance decrease. Throughout all scrambling conditions, the order of the models concerning their performance stays the same. This means that in all conditions, Fiber performed the best, MDETR the worst, and the three GLIP models in between. All the described effects can be seen in Figure 2a. For different evaluation metrics (e.g., Recall@5 as seen in Appendix C.3), the influence of syntactic scrambling is even less accentuated.

Looking at the relative robustness of the models directly (Figure 2b), we can see that the performance of the models directly predicts their robustness. For example, Fiber is not only the model with the highest accuracy but also the model with the lowest drop in accuracy when applied to scrambled sentences.

The models' performance depends not only on the level of scrambling but also on the category of the phrase on which the model is evaluated. In Figure 2c, we compared the influence of phrase category on model robustness. For some categories (e.g., animals), all models are robust against scrambling. For other categories, on the other hand (e.g., clothes), not all models are that robust against scrambling. In these cases, the robustness of the models does not vary much. However, there are also categories (e.g., instruments) with no general trend in the data, and robustness varies significantly between models.

5.3. Discussion

Our results indicate that even though word order does influence the performance of state-of-the-art models, those models are surprisingly robust against syntactic manipulations. This observation suggests that the evaluated phrase grounding models exhibit only a minimal reliance on linguistic structure. Our findings in the domain of phrase grounding align with results from other areas.

[1] https://paperswithcode.com/sota/phrase-grounding-on-flickr30k-entities-test, last accessed: 19.06.24

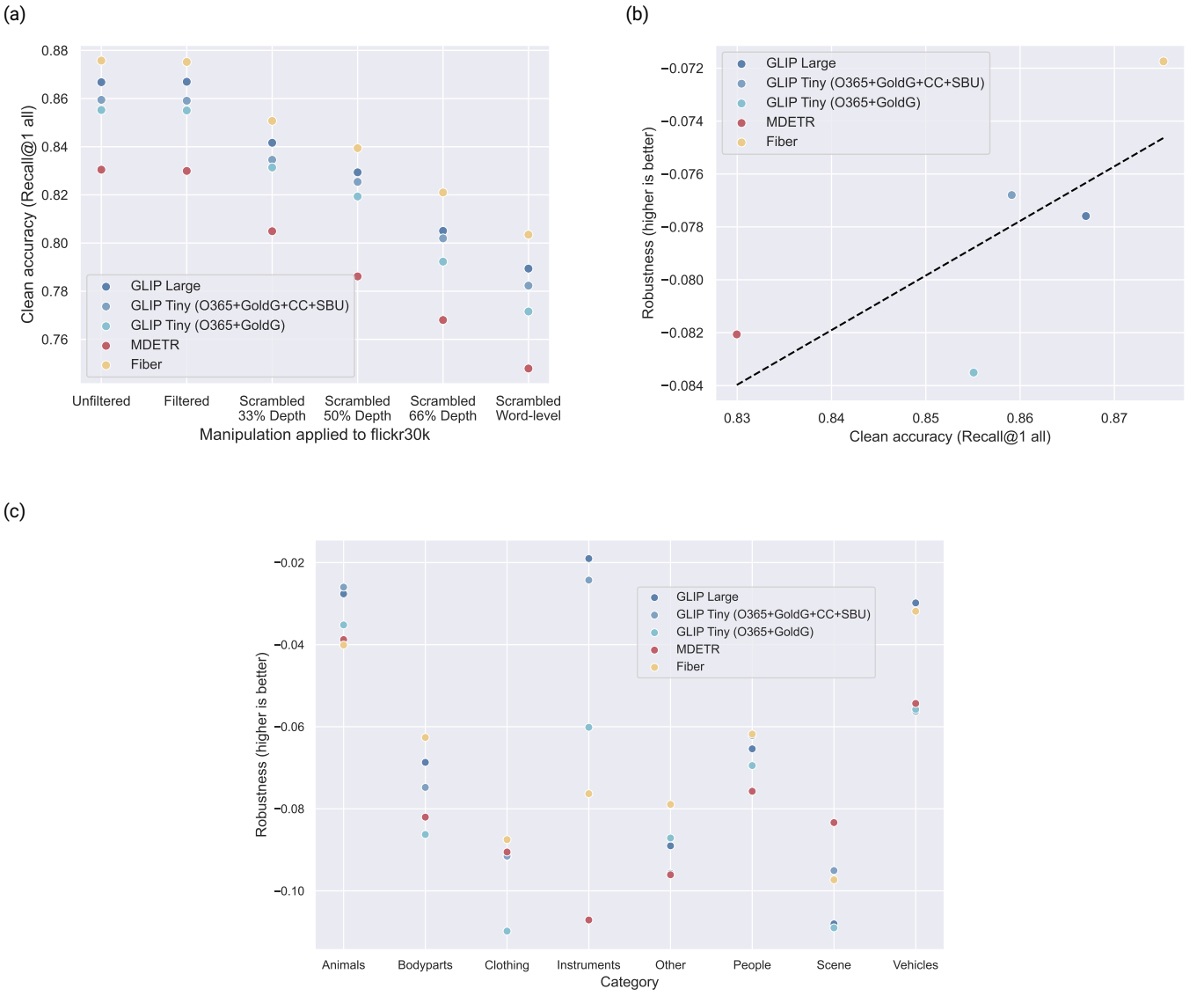

Figure 2: **Overall performance and robustness of models on manipulated data. (a) Influence of different levels of scrambling on model performance.** Results on the unfiltered dataset describe the recall of the models on the original flickrf30k entities test dataset. The filtered dataset contains only sentences with a semantic tree depth greater than two (see section 4). **(b) Relationship between performance and relative robustness.** Circles indicate different models, while the dotted black line indicates the best-fit regression. Relative robustness is measured as the difference between the model's performance on the filtered dataset and the dataset where each sentence was scrambled at the word level. **(c) Influence of phrase category on relative robustness to scrambling.** Relative robustness is measured as the difference between the model's performance on the filtered dataset and the dataset where each sentence was scrambled at the word level.

For instance, in the grounding of visual referring expressions, word order appears to have limited influence on model performance [1, 4]. Similarly, in unimodal natural language understanding and question-answering tasks, it has been shown that state-of-the-art models do not heavily depend on syntactic structure [19, 30, 8]. Together, these findings suggest that state-of-the-art models engaged in natural language processing tasks often bypass traditional linguistic structures, using alternative mechanisms to achieve high performance. This could imply that the statistical patterns and contextual embeddings primarily drive such models in the data rather than by explicit syntactic rules.

Our analysis revealed the following pattern: models with higher baseline performance also displayed greater robustness to syntactic scrambling. This correlation was consistent across all levels of scrambling and all models tested. This discovery prompts essential questions about the nature of the learned representations in these models. It suggests that the features or patterns contributing to high performance also play a role in their robustness against syntactic perturbations. Therefore, the high performance of the models

does not necessarily imply an accurate syntactic representation and understanding.

Upon examining the robustness of various phrase categories, we observed interesting differences. For instance, the animal category exhibited high robustness across all models, while categories like clothing showed lower robustness. This suggests that the models may rely differently on syntactic structure based on the semantic content being conveyed. The highly robust categories may be those where visual features play a particularly significant role or where the relationship between the textual description and the visual representation is more direct. On the other hand, categories with lower robustness may involve more intricate or abstract relationships between text and image, possibly necessitating a greater reliance on syntactic structure.

6. Conclusions and Limitations

While our study provides valuable insights, it is important to acknowledge its limitations. Our experiments were conducted exclusively on the flickr30k Entities dataset, which may limit the generalizability of our findings due to dataset-specific biases. Although the syntactic scrambling method is informative, it may not capture all linguistic corruptions or real-world variations. Our model selection was restricted to those with publicly available weights and using MDETR annotations, potentially not fully representing the spectrum of phrase grounding models. We recognize that the focus on Recall@k metrics may have overlooked other performance measures, and the absence of human performance comparisons on scrambled sentences limits our reference points. Additionally, our study primarily emphasizes input-output relationships without exploring internal model mechanics.

In order to overcome these limitations, future research should broaden the scope of datasets and investigate a wider range of syntactic manipulations. For example, implementing advanced methodologies similar to those described in [1], where human crowd workers intentionally scrambled sentences to deceive models, could offer deeper insights. Encompassing a more diverse set of models, integrating comparisons with human performance, and conducting a more comprehensive analysis of model internals could further clarify how these models attain robustness against syntactic manipulations. Lastly, a more thorough examination of the impact of specific linguistic structures on model performance could generate valuable insights for enhancing phrase grounding systems.

In conclusion, this study offers valuable insights into the resilience of state-of-the-art phrase grounding models when subjected to syntactic manipulations. Our findings indicate that although these models demonstrate considerable resilience to changes in linguistic structure, their performance is still somewhat affected by word order. Notably, we observed a link between a model's initial performance and its ability to withstand syntactic scrambling, suggesting that high-performing models may utilize alternative mechanisms beyond linguistic structures. These findings have important implications for our comprehension of how multimodal language models process and interpret information, underscoring the necessity for further investigation into the nature of learned representations in these systems.

References

[1] A. R. Akula, S. Gella, Y. Al-Onaizan, S. Zhu, and S. Reddy. Words aren't enough, their order matters: On the robustness of grounding visual referring expressions. In D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6555–6565. Association for Computational Linguistics, 2020.

[2] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 213–229. Springer, 2020.

[3] H. Chen, H. Zhang, P. Chen, J. Yi, and C. Hsieh. Attacking visual language grounding with adversarial examples: A case study on neural image captioning. In I. Gurevych and Y. Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2587–2597. Association for Computational Linguistics, 2018.

[4] V. Cirik, L.-P. Morency, and T. Berg-Kirkpatrick. Visual Referring Expression Recognition: What Do Systems Actually Learn?, May 2018.

[5] Z. Dou, A. Kamath, Z. Gan, P. Zhang, J. Wang, L. Li, Z. Liu, C. Liu, Y. LeCun, N. Peng, J. Gao, and L. Wang. Coarse-to-fine vision-language pre-training with fusion in the backbone. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

[7] D. Hendrycks and T. G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[8] R. Jia and P. Liang. Adversarial Examples for Evaluating Reading Comprehension Systems, July 2017.

[9] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion. MDETR - modulated detection for end-to-end multi-modal understanding. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1760–1770. IEEE, 2021.

[10] N. Kitaev, S. Cao, and D. Klein. Multilingual constituency parsing with self-attention and pre-training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505, Florence, Italy, July 2019. Association for Computational Linguistics.

[11] N. Kitaev and D. Klein. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[12] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and F.-F. Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations, 2016.

[13] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J. Hwang, K. Chang, and J. Gao. Grounded language-image pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10955–10965. IEEE, 2022.

[14] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár. Microsoft coco: Common objects in context, 2015.

[15] C. Liu, Y. Dong, W. Xiang, X. Yang, H. Su, J. Zhu, Y. Chen, Y. He, H. Xue, and S. Zheng. A comprehensive study on robustness of image classification models: Benchmarking and rethinking. *CoRR*, abs/2302.14301, 2023.

[16] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[17] V. Ordonez, G. Kulkarni, and T. Berg. Im2text: Describing images using 1 million captioned photographs. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

[18] R. Peri, S. M. Jayanthi, S. Ronanki, A. Bhatia, K. Mundnich, S. Dingliwal, N. Das, Z. Hou, G. Huybrechts, S. Vishnubhotla, D. Garcia-Romero, S. Srinivasan, K. J. Han, and K. Kirchhoff. Speechguard: Exploring the adversarial robustness of multimodal large language models. *CoRR*, abs/2405.08317, 2024.

[19] T. M. Pham, T. Bui, L. Mai, and A. Nguyen. Out of order: How important is the sequential order of words in a sentence in natural language understanding tasks? In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 1145–1160. Association for Computational Linguistics, 2021.

[20] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *Int. J. Comput. Vis.*, 123(1):74–93, 2017.

[21] J. Qiu, Y. Zhu, X. Shi, F. Wenzel, Z. Tang, D. Zhao, B. Li, and M. Li. Benchmarking robustness of multimodal image-text models under distribution shift. 2022.

[22] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

[23] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generation. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR, 2021.

[24] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015.

[25] C. Schlarmann and M. Hein. On the adversarial robustness of multi-modal foundation models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023 - Workshops, Paris, France, October 2-6, 2023*, pages 3679–3687. IEEE, 2023.

[26] M. Shah, X. Chen, M. Rohrbach, and D. Parikh. Cycle-consistency for robust visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6649–6658. Computer Vision Foundation / IEEE, 2019.

[27] S. Shao, Z. Li, T. Zhang, C. Peng, G. Yu, X. Zhang, J. Li, and J. Sun. Objects365: A large-scale, high-quality dataset for object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8429–8438, 2019.

[28] P. Sharma, N. Ding, S. Goodman, and R. Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. 2018.

[29] C. Si, S. Wang, M. Kan, and J. Jiang. What does BERT learn from multiple-choice reading comprehension datasets? *CoRR*, abs/1910.12391, 2019.

[30] K. Sinha, R. Jia, D. Hupkes, J. Pineau, A. Williams, and D. Kiela. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In M. Moens, X. Huang, L. Specia, and S. W. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 2888–2913. Association for Computational Linguistics, 2021.

[31] S. Sugawara, P. Stenetorp, K. Inui, and A. Aizawa. Assessing the benchmarking capacity of machine reading comprehension datasets. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8918–8927. AAAI Press, 2020.

[32] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In Y. Bengio and Y. LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[33] A. Talman, M. Apidianaki, S. Chatzikyriakidis, and J. Tiedemann. How does data corruption affect natural language understanding models? A study on GLUE datasets. In V. Nastase, E. Pavlick, M. T. Pilehvar, J. Camacho-Collados, and A. Raganato, editors, *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics, *SEM@NAACL-HLT 2022, Seattle, WA, USA, July 14-15, 2022*, pages 226–233. Association for Computational Linguistics, 2022.

[34] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[35] K. Yang, W. Lin, M. Barman, F. Condessa, and J. Z. Kolter. Defending multimodal fusion models against single-source adversaries. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 3340–3349. Computer Vision Foundation / IEEE, 2021.

[36] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguistics*, 2:67–78, 2014.

[37] J. Zhang, T. Pang, C. Du, Y. Ren, B. Li, and M. Lin. Benchmarking large multimodal models against common corruptions. *CoRR*, abs/2401.11943, 2024.

A. GitHub Project Info

Link to GitHub repositories:

Public repository (does not contain used models):
https://github.com/lreining/robustness_of_phrase_grounding_models
Private repository (does contain all used models - ask us for access):
https://github.com/finnradatz/robustness_of_phrase_grounding_models

List of GitHub Ids:

- Hannah Willkomm: `hannah-sophie`
- Doğukan Bağcı: `DogukanBG`
- Finn Radatz: `finnradatz`
- Lars Reining: `lreining`
- Tim-Michael Krieg: `TimM4ster`

Please note that we also collaborated with Lightning.ai. As a result, the individual code contributions to get the models running are not fully reflected in the commit history. For a detailed description of each person's implementations, please refer to the following section.

B. Individual Contributions

All members of our group provided valuable contributions throughout the project. Everyone participated in formulating the project report and both the project pitch and final presentation. In terms of the implementation of our experiments, we declared two sub-groups. One group was tasked with the natural language pre-processing (i.e. scrambling) and analysis of the data, while the other group was supposed to get the models to run smoothly. This included modifications to the original code of the individual models, such that our newly created input is passed and evaluated correctly. Generally, we all attended to joint meetings regularly in order to keep ourselves updated on the process of the other members.

Hannah Willkomm Hannah Willkomm provided valuable knowledge for the data's preprocessing and, as part of the NLP group, the necessary code to scramble the sentences and visualize the syntax trees of the scrambled sentences.

Doğukan Bağcı Doğukan Bağcı provided helpful research material and, as part of the model group, the necessary code to run the MDETR model. He also engaged in analyzing our data.

Finn Radatz Finn Radatz provided research background and, as part of the model group, provided the necessary code to run the different GLIP models. He also participated in the project pitch to present our initial idea and presented a summary of our findings in the final presentation.

Lars Reining As part of the NLP group, Lars Reining provided important code for the implementation of the word and sentence scrambling. Further, he engaged in the analysis of our data and provided plots for the final analyses.

Tim Krieg Tim Krieg, as part of the model subgroup, provided the necessary code to run the FIBER model. He also participated in the project pitch to present our initial idea and presented a summary of our findings in the final presentation.

C. Additional Content

C.1. Damerau–Levenshtein Distances for Manipulations

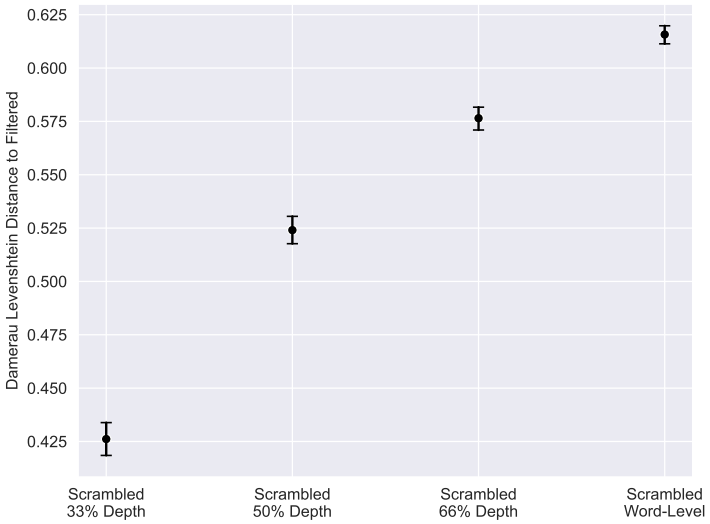

Figure 3: **Damerau–Levenshtein distance for the different manipulation of the sentences** Damerau–Levenshtein distance demonstrates that higher percentages used in our corruption method lead to less preservation of the original sentence structure as those result in higher distance values.

C.2. Models

In the following, we briefly present the models included in our analysis.

GLIP [13] We included three different models of the GLIP model family. The first one is GLIP-T (C) (GLIP Tiny (O365+GoldG)), which is built on the Swin-Tiny backbone and is pre-trained on O365 and GoldG. It utilizes 0.8 million human-annotated gold grounding data curated by MDETR, which includes datasets such as flickr30K, VG Caption, and GQA. The second one is GLIP-T (GLIP Tiny (O365+GoldG+CC+SBU)), which is also based on the Swin-Tiny backbone and is pre-trained on the following data: O365, GoldG as in GLIP-T (C), and Cap4M, which comprises 4 million image-text pairs collected from the web with boxes generated by GLIP-T (C). The third one is GLIP-L (GLIP Large), which is built on Swin-Large and is trained with FourODs (2.66 million data), 4 detection datasets including Objects365, OpenImages, Visual Genome (excluding COCO images), and ImageNetBoxes. It also includes GoldG as in GLIP-T (C), and CC12M+SBU, which encompasses 24 million image-text data collected from the web with generated boxes.

FIBER [5] FIBER (**F**usion-**I**n-the-**B**ackbone-based transform**ER**) is the first vision-language model to combine a high- (image-text retrieval, VQA, etc.) and region-level (phrase grounding, etc.) understanding of images. This is accomplished by introducing a cross-attention mechanism within the SWIN-based architecture, enabling cross-modal interactions. Additionally, FIBER also makes use of a two-stage training process, where *coarse*-grained pre-training was done on image-text and *fine*-grained pre-training on image-text-box data.
For our experiments, we used the fine-grained pre-trained version of FIBER. This version was first *coarsly* pre-trained on COCO [14], VG [12], SBU [17] and GCC [28] and subsequently *fine-tuned* on the flickr30k [20], MixedNoCOCO [9] and Objects365 (O365) [27] datasets.

MDETR [13] This is a multimodal detection model which is based on the DETR detector [2]. Just like the DETR detector, MDETR uses a convolutional neural network for encoding the image. Additionally, a pre-trained transformer language model for encoding the text inputs is added to the architecture. The extracted features are then projected in a shared embedding space which are then processed by transformer encoder, named the cross encoder. In the last step, a transformer decoder is applied on the output of the cross encoder in order to generate predictions. In our analysis we use the MDETR model with a ResNet-101 backbone [6].

C.3. Results for Recall@5

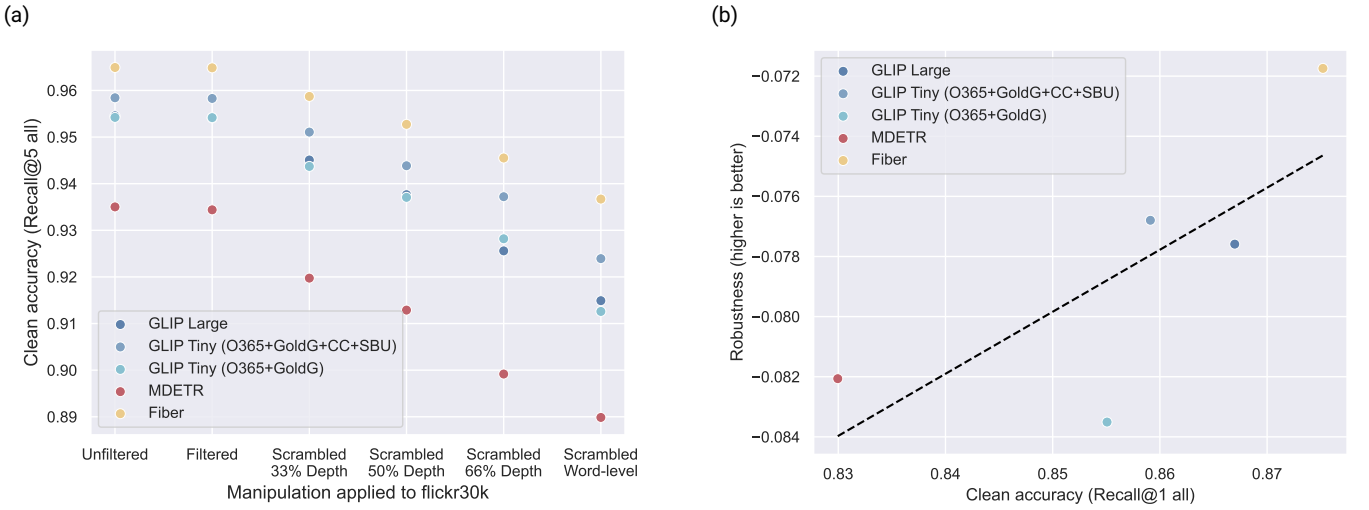

Figure 4: **Overall performance and robustness of models on manipulated data as measured by Recall@5. (a) Influence of different levels of scrambling on model performance.** Results on the unfiltered dataset describe the recall of the models on the original flickr30k entities test dataset. The filtered dataset contains only sentences with a syntactic tree depth of greater than two (see section 4). **(b) Relationship between performance and relative robustness.** Circles indicate different models while the dotted black line indicates the best fit regression. Relative robustness is measured as the difference between the models performance on the filtered dataset and the dataset where each sentence was scrambled at word-level.