Harlequin: Color-driven Generation of Synthetic Data for Referring Expression Comprehension

Luca Parolari Elena Izzo Lamberto Ballan

Department of Mathematics "Tullio Levi-Civita", University of Padova, Italy

Abstract

Referring Expression Comprehension (REC) aims to identify a particular object in a scene by a natural language expression, and is an important topic in visual language understanding. State-of-the-art methods for this task are based on deep learning, which generally requires expensive and manually labeled annotations. Some works tackle the problem with limited-supervision learning or by relying on Large Vision and Language models. However, the development of techniques to synthesize labeled data is overlooked. In this paper, we propose a novel pipeline that generates artificial data for the REC task, taking into account both textual and visual modalities. The pipeline processes existing data to create variations in the annotations. Then, it generates an image using altered annotations as guidance. The result of this pipeline is a new dataset, termed Harlequin, made by more than 1M queries. This approach eliminates manual data collection and annotation, enabling scalability and facilitating arbitrary complexity. We pre-train two REC models on Harlequin, then fine-tuned and evaluated on human-annotated datasets. Our experiments show that the pre-training on artificial data is beneficial for performance.

1. Introduction

The research progress in Referring Expression Comprehension task has been made possible thanks to an active development of datasets. Since 2015, Flickr30k Entities [9], ReferIt [6], RefCOCO, and two variants RefCOCO+ and RefCOCOg, [8, 15] were released. These datasets are all human-labeled and consist of triplets composed of an image, a referring expression, and a bounding box. However, the gathering and annotation of such data is time-consuming and resource-intensive, representing a critical bottleneck for the collection of sufficiently large training sets.

Current works face this issue exploring limited supervision learning techniques [4, 10, 13] or rely on large Vision and Language models pre-trained on a massive amount of multimodal data [5]. However, the development of techniques and pipelines to create new, reasoning-oriented datasets is overlooked since limited by fine-grained annotations required by the Referring Expression Comprehension task. Some works explore the generation of the queries by either working on their properties or structure [1, 3, 12]. However a method to generate both queries and images has not been investigated yet.

In this paper, we propose a pipeline for generating synthetic data for the Referring Expression Comprehension task, taking into account both textual and visual modalities. Thanks to advancements in generative AI, we argue that (i) the process of manual collection and annotation of data for this task can finally be avoided, and (ii) new benchmarks with arbitrary size and complexity can be created. To verify those hypothesis we generated a new artificial dataset, named Harlequin, consisting of train, validation, and test sets. Harlequin is the first dataset totally synthetic generated for the Referring Expression Comprehension task, and the experiments show that its use in pre-training stage can boost the results on real data. Moreover, we illustrate its benefit in reducing labeling effort, and errors in annotations.

Our contributions can be summarized as follows: (i) We propose a novel pipeline for generating synthetic data, reducing to zero the human effort for annotations collection; (ii) We introduce Harlequin, a new dataset for Referring Expression Comprehension task totally synthetic generated; (iii) We analyse the generated dataset both qualitatively and quantitative proving its effectiveness if used in a pre-training stage to transfer knowledge on real datasets.

2. The proposed pipeline

The proposed approach, depicted in Fig. 1, relies on two components to generate synthetic data for Referring Expression Comprehension. The former, termed **Annotation Generation Engine**, is in charge of creating new annotations by varying attributes in the referring expression to guide the image generation. The latter, named **Image Generation Engine**, is responsible to synthesize images enforcing the

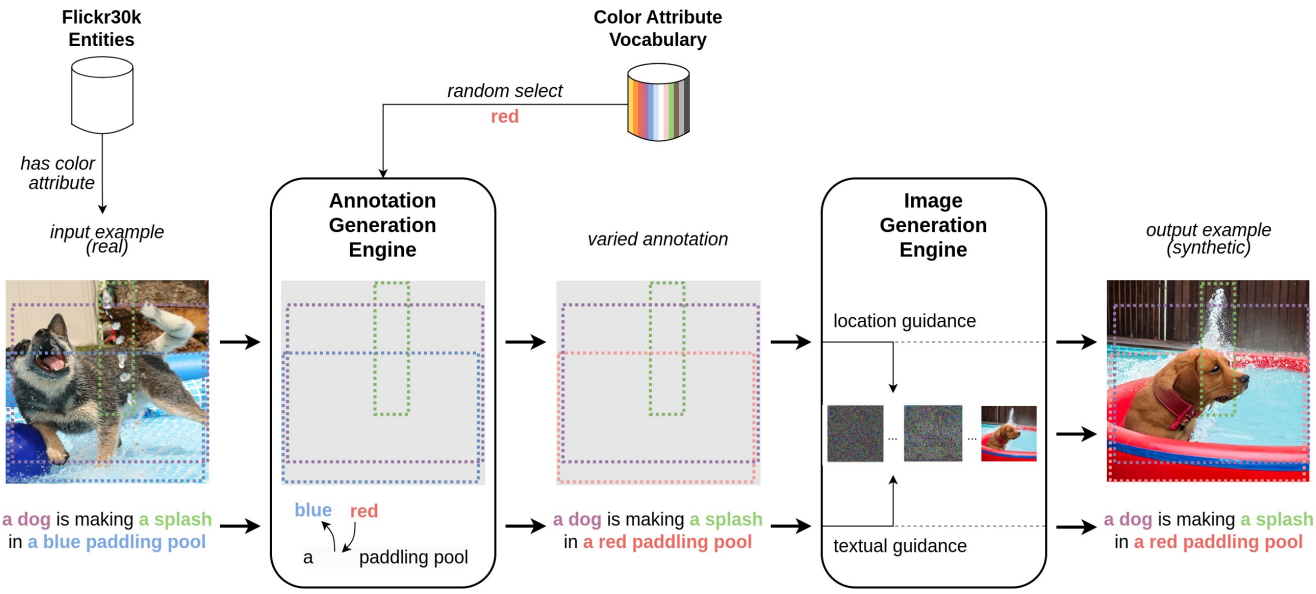

Figure 1. Our pipeline. It processes existing samples from Flickr30k Entities data. We select the ones characterized by at least one *color* attribute in their referring expressions. The Annotation Generation Engine processes sample's caption, referring expressions and locations where the color attribute is replaced with a randomly chosen color. The caption is updated accordingly. Then, the Image Generation Engine create the new image using new annotations provided by the Annotation Generation Engine as guidance for the generation.

guidance provided by the Annotation Generation Engine.

The pipeline processes existing sample to generate new, synthetic data. We feed Flickr30k Entities samples to the pipeline because every image is annotated with a sentence, yielding many referring expressions. From a generative point of view, this setting alleviates the amount of guessing and constrains the possible space of images that can be generated to a subset, where objects are precisely described and spatially located.

The alteration we chose to apply on existing data regards the attributes, specifically the color attribute. The reasons behind this choice are manifold. In particular: *(i)* it is a straightforward variation of the textual content but the impact on the visual modality is intuitive and relevant; *(ii)* its variation does not affect the position of the object in the image, retaining the realistic layout of objects in the image; (iii) it is not ambiguous and easily applicable on a synthetic image; *(iv)* it is the most used attribute in Flickr30k Entities.

2.1. The Annotation Generation Engine

The Annotation Generation Engine (AGE) is a function defined over the set of annotations $\mathcal{A}$. It is specifically designed for Referring Expression Comprehension task and produces compatible annotations by altering existing samples: $\phi : \mathcal{A} \rightarrow \mathcal{A}$. The AGE component takes an annotation $\boldsymbol{a} = (\boldsymbol{c}, E)$ in input. It consists in a caption $\boldsymbol{c} = [c_1, \cdots, c_L]$ of L tokens and a non empty set of entities $E = \{(\boldsymbol{q}_i, \boldsymbol{l}_i)\}_{i=1}^{N}$. Each entity is described by the textual form of a referring expression $\boldsymbol{q_i} = [c_j, \ldots, c_k]$ with $1 \leq j \leq k \leq L$ from a subset of contiguous tokens in $\boldsymbol{c}$, while $\boldsymbol{l_i} = [\alpha_{\text{min}}, \beta_{\text{min}}, \alpha_{\text{max}}, \beta_{\text{max}}]$ is with top-left and bottom-right coordinates of the referred object and the location of the referred object. The AGE returns a new annotation where the p-th referring expression is varied by replacing a color attribute, $p \in [1, N]$. The location is not altered. Tokens in the caption are updated accordingly to the new referring expression, while other referred objects are not varied and serve as context. Mathematically, the output of $\phi(\boldsymbol{a})$ is $\hat{\boldsymbol{a}} = (\hat{\boldsymbol{c}}, \hat{E})$ where $\hat{\boldsymbol{c}} = [c_1, \cdots, c_{j-1}, \hat{c}_j, \cdots, \hat{c}_k, \cdots c_L]$ and $\hat{E} = \{\hat{\boldsymbol{q}}_p, \boldsymbol{l}_p\} \cup \{(\boldsymbol{q}_i, \boldsymbol{l}_i)\}_{i=1, i\neq p}^{N}$, with $\hat{\boldsymbol{q}}_p = [\hat{c}_j, \cdots, \hat{c}_k]$ the new referring expression where the color attribute is changed. Specifically, we replace in $\boldsymbol{q}_p$ the color with a new randomly sampled one. Sampling is done on a vocabulary C of 12 color attributes: black, gray, white, red, orange, yellow, green, cyan, blue, purple, pink, brown [16]. The variation function ϕ is applied 6 times ($|C|/2$) per referring expression with color attribute. We chose 6 as a trade-off between the number of annotations generated and the variability introduced through multiple sampling.

The current definition of ϕ keeps fixed all objects' locations and $N - 1$ referring expressions. This is done to preserve the spatial arrangement of the objects, i.e. the layout, and the image context. Objects' locations are particularly relevant as they express complex semantic meaning. In order to keep this rich semantic, in this work we prefer to focus on variation of text which is more intuitive both to generate and evaluate.

2.2. The Image Generation Engine

The Image Generation Engine (IGE) is responsible for generating synthetic images. This component receives an annotation $\hat{\boldsymbol{a}}$ obtained from the Annotation Generation Engine. It returns an image $I \in \mathcal{I}$ from the domain of images encoding semantic information expressed in $\hat{\boldsymbol{a}}$. More in details, we define the IGE as a function $\psi : \mathcal{A} \rightarrow \mathcal{I}$: $\psi(\hat{\boldsymbol{a}}) = I$. We implement the Image Generation Engine component with Grounded-Language-to-Image Generation (GLIGEN) [7]. GLIGEN is a generative model based on Stable Diffusion [11] that allows fine-grained control over the output image. It generates an image representing its content through a caption as in Stable Diffusion, but it attends also to a set of pairs (referring expression, object location), namely entities. These entities instruct GLIGEN with the location of objects and how to depict them. The more accurate the positioning of objects and fidelity to descriptions, the better the supervision signal for the Referring Expression Comprehension task.

3. Harlequin dataset

We introduce the first totally synthetic generated dataset for the Referring Expression Comprehension task, termed Harlequin,[1] collected via our pipeline. The dataset is originated from Flickr30k Entities: we select samples characterized by referring expressions containing the *color* attribute to variate them. Since the Image Generation Engine is eager in terms of resources, we first generate all the new annotations with Annotation Generation Engine using the selected samples as seeds. Secondly, we parallelized the image synthesis over 4 NVIDIA RTX A6000 GPUs. We adopt a frozen instance of GLIGEN in "generation" mode with "text + box" modality using code and weights provided by authors,[2] and setting the batch size to 1. The generation of the whole dataset took 15 days per GPU.

Harlequin, depicted in Fig. 2, comprises a total of 286,948 synthetic images and 1,093,181 annotations targeting color attributes. It has 2.60±1.14 words per referring expression on average, in line with Flickr30k Entities statistics. The median value is 2, while the longest referring expression is 14 words. Harlequin follows Flickr30k Entities' data splits. It provides 988,342 annotations over 259,930 images for the training set, 52,554 annotations over 13,584 images for the validation set, and 52,284 annotations over 13,434 images for the test set. Harlequin doubles the amount of referring expression in Flickr30k, the largest dataset available in the literature, and provides a noticeably larger amounts of images.

The generated images display the same objects under various orientations and on different backgrounds, increasing the variability and complexity of Harlequin with respect to Flickr30k Entities, while retaining its supervision signal. Moreover, we bring up that the used variation function ϕ inevitably leads to the generation of unrealistic-colored objects (e.g. "the blue dog"). Independently of that, the results show that Referring Expression Comprehension models learn a robust representation from Harlequin. This is coherent with the fact that humans are usually capable of identifying an object regardless of its color and use this information to disambiguate similar objects.

Figure 2. Examples produced by our pipeline. On the top, we show reference images along with their annotations from Flickr30k Entities. On the bottom, we report some generated variations. Colors are altered and guide, along with objects' locations, the image synthesis.

4. Experiments

Results We validate our synthetic dataset, Harlequin, by pre-training state-of-the-art models on it and fine-tuning them on realistic datasets. The Tab. 1 reports fine-tuning results of TransVG [2] and VLTVG [14] on three realistic

[1] Harlequin, or Arlecchino in Italian, is a character from the Italian commedia dell'arte known for his colorful patched costume.

[2] https://github.com/gligen/GLIGEN

Method	RefCOCO			RefCOCO+			RefCOCOg	
	val	testA	testB	val	testA	testB	val	test
TransVG [2]:								
R	63.33	69.05	55.62	64.69	69.02	**55.76**	64.04	63.22
S→R	**65.77**	**70.66**	**56.80**	**66.66**	**72.01**	55.66	**65.13**	**64.33**
(Impr.)	+2.44	+1.61	+1.18	+1.97	+2.99	-0.10	+1.09	+1.11
VLTVG [14]:								
R	**69.66**	74.33	**61.35**	70.83	76.02	**61.71**	**70.57**	**70.03**
S→R	69.60	**75.76**	61.14	**71.46**	**77.16**	61.30	70.04	69.57
(Impr.)	-0.06	+1.43	-0.21	+0.63	+1.12	-0.41	-0.53	-0.46

Table 1. Results. We show the performance of two methods, TransVG and VLTVG, on the Referring Expression Comprehension task with pre-training on Harlequin (*Synth→Real* denoted by *S→R*) and without (*Real* denoted by *R*). *Impr.* denotes improvement. The pre-training shows superior or comparable performance on three benchmarks: RefCOCO, RefCOCO+ and RefCOCOg. We report the accuracy in percentage.

Test set	TransVG [2]		VLTVG [14]	
	Real	Synth→Real	Real	Synth→Real
RC val*	64.35	69.50 (+5.15)	77.15	77.11 (-0.04)
RC testA*	68.73	72.91 (+4.18)	79.46	81.48 (+2.02)
RC testB*	56.46	60.11 (+3.65)	68.95	70.28 (+1.33)
RC+ val*	68.91	70.07 (+1.16)	75.79	77.13 (+1.34)
RC+ testA*	71.63	74.20 (+2.59)	79.13	80.95 (+1.82)
RC+ testB*	55.73	57.33 (+1.60)	65.65	63.82 (-1.83)
RCg val*	62.37	65.04 (+2.67)	73.23	73.09 (-0.14)
RCg test*	62.12	64.94 (+2.82)	73.05	73.70 (+0.65)

Table 2. Ablation study. We evaluate the performance of TransVG and VLTVG on a subset of test sets where referring expressions contains at least one color attribute. Resulting test sets, marked by *, have different amount of annotations with respect to original test sets: 23.2, 37.5, 17.8% for RefCOCO (RC), 34.6, 37.5, 26.8% for RefCOCO+ (RC+) and 41.4, 41.7% for RefCOCOg (RCg). Columns *Real* and *Synth→Real* show the performance without or with pre-training on Harlequin. We report accuracy in percentage.

datasets: RefCOCO, RefCOCO+ and RefCOCOg. Moreover, it reports the results obtained by training the models directly on realistic benchmarks as baselines. The performance is evaluated via standard accuracy on the test sets.

TransVG shows homogeneous improvement among all datasets, gaining up to 2.44%, 1.61%, and 1.18% in RefCOCO splits. As concerns VLTVG, despite its higher baseline performance with respect to TransVG, it shows a remarkable 1.43% and 1.12% improvement on RefCOCO and RefCOCO+'s testA. We recall that there is no overlap between the pre-training data, synthetically generated from Flickr30k Entities, and the fine-tuning datasets. The reported improvement emerges in a cross-dataset setting.

Results are encouraging: they demonstrate that pre-training on synthetically generated data is feasible also in the Referring Expression Comprehension task. Annotations required by this task challenge generative models, where their artistic traits need to deal with fine-grained constraints on objects' locations and descriptions. Nevertheless, our pipeline proves that the generation and collection of heavily annotated data with zero human effort is possible. This is an important milestone and opens a wide range of future directions where data can be crafted to overcome the increasing need for annotations.

Impact of the Color Attribute In this section we evaluate the impact of our pre-training on realistic samples with *color* attribute. We follow the same training scheme described previously. However, here the test sets are limited to samples whose referring expressions have at least one color. As shown in Tab. 2, TransVG [2] demonstrates a boost in performance among RefCOCO family datasets, with remarkable $+5.15\%$, $+4.18\%$ and $+3.65\%$ on RefCOCO. The pre-training shows superior or comparable performance also for VLTVG [14], with the except of testB for RefCOCO+. We recall that no changes to model's architecture have been made to encode extra knowledge about colors. The improvement is solely guided by learning patterns from data.

However, these results were expected. As a matter of fact, Harlequin is mainly composed by referring expressions which contain a color attribute. Consequently, models pre-trained on our dataset primarily acquire generalization capabilities in identifying and distinguishing objects with different colors. Moreover, these results suggest that increasing the variability and amount of data for a specific attribute impacts the model's performance on real data with the same attribute. Thus, we believe that synthetic data can further boost results on real data by extending our method on other attributes.

5. Conclusion

In this work, we introduce Harlequin, the first synthetic dataset for the Referring Expression Comprehension task, collected via a novel pipeline. Our strategy manipulates color attributes in existing referring expressions to generate datasets with arbitrary size and complexity without human effort. We validated our method by improving the state-of-the-art results on real data after a pre-training on Harlequin.

In future work, we plan to investigate the potential and flexibility of our pipeline, extending it on other attributes besides the color. Second, we aim to automatize the generation of the referring expressions via prompting strategies.

References

[1] Zhenfang Chen, Peng Wang, Lin Ma, Kwan-Yee K. Wong, and Qi Wu. Cops-ref: A new dataset and task on compositional referring expression comprehension. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10083–10092. Computer Vision Foundation / IEEE, 2020. 1

[2] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1749–1759. IEEE, 2021. 3, 4

[3] Haojun Jiang, Yuanze Lin, Dongchen Han, Shiji Song, and Gao Huang. Pseudo-q: Generating pseudo language queries for visual grounding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15492–15502. IEEE, 2022. 1

[4] Jianglin Jin, Jiabo Ye, Xin Lin, and Liang He. Pseudo-query generation for semi-supervised visual grounding with knowledge distillation. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE, 2023. 1

[5] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas 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. 1

[6] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 787–798. ACL, 2014. 1

[7] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. GLIGEN: open-set grounded text-to-image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 22511–22521. IEEE, 2023. 3

[8] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 11–20. IEEE Computer Society, 2016. 1

[9] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2641–2649. IEEE Computer Society, 2015. 1

[10] Davide Rigoni, Luca Parolari, Luciano Serafini, Alessandro Sperduti, and Lamberto Ballan. Weakly-supervised visual-textual grounding with semantic prior refinement. In *34th British Machine Vision Conference 2022, BMVC 2022, Aberdeen, UK, November 20-24, 2023*, page 229. BMVA Press, 2023. 1

[11] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE, 2022. 3

[12] Mikihiro Tanaka, Takayuki Itamochi, Kenichi Narioka, Ikuro Sato, Yoshitaka Ushiku, and Tatsuya Harada. Generating easy-to-understand referring expressions for target identifications. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 5793–5802. IEEE, 2019. 1

[13] Josiah Wang and Lucia Specia. Phrase localization without paired training examples. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4662–4671. IEEE, 2019. 1

[14] Li Yang, Yan Xu, Chunfeng Yuan, Wei Liu, Bing Li, and Weiming Hu. Improving visual grounding with visual-linguistic verification and iterative reasoning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 9489–9498. IEEE, 2022. 3, 4

[15] 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*, pages 69–85. Springer, 2016. 1

[16] Yang Zhan, Zhitong Xiong, and Yuan Yuan. RSVG: exploring data and models for visual grounding on remote sensing data. *IEEE Trans. Geosci. Remote. Sens.*, 61:1–13, 2023. 2