Grounding Plural Phrases: Countering Evaluation Biases by Individuation

Julia Suter **Letitia Parcalabescu** **Anette Frank**
Department of Computational Linguistics
Heidelberg University, 69120 Heidelberg
{suter,parcalabescu,frank}@cl.uni-heidelberg.de

Abstract

Phrase grounding (PG) is a multimodal task that grounds language in images. PG systems are evaluated on well-known benchmarks, using *Intersection over Union (IoU)* as evaluation metric. This work highlights a disconcerting bias in the evaluation of grounded *plural phrases*, which arises from representing *sets of objects* as a *union box* covering all component bounding boxes, in conjunction with the IoU metric. We detect, analyze and quantify an *evaluation bias* in the grounding of plural phrases and define a *novel metric, c-IoU,* based on a union box's component boxes. We experimentally show that our new metric greatly alleviates this bias and recommend using it for fairer evaluation of plural phrases in PG tasks.

1 Introduction

Phrase grounding (PG) describes the multimodal task of identifying objects in images and connecting them to free-form phrases in a textual description (caption). A phrase usually describes one, or sometimes several, specific objects.

Grounding phrases in image regions provides an essential link between texts and images and serves as a foundation for multimodal understanding tasks, including sentence-to-image alignment, Visual QA, Visual Common-sense Reasoning (VCR), etc.

Benchmarks for training and evaluating PG systems (Everingham et al., 2010; Lin et al., 2014; Kazemzadeh et al., 2014; Plummer et al., 2015; Krishna et al., 2017) generally provide rectangular bounding boxes as ground truth (GT). Therefore a PG ground truth is represented as a phrase linking to a (gold) bounding box enclosing the image patch referred to by the phrase. Some datasets provide pixel segmentation masks (Lin et al., 2014), which enable more precise evaluations but are more difficult and costly to produce. Thus, the trend towards annotating bounding boxes persists in recent datasets (Ilinykh et al., 2019).

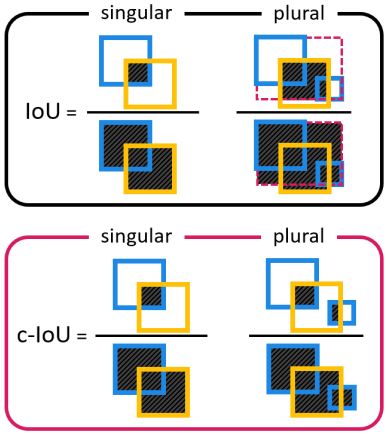

Figure 1: Illustration of how to compute the evaluation metric *IoU* and its adaptation *c-IoU* (with ground truth (GT) bounding boxes in blue, and predicted boxes in yellow). The numerator represents the computed intersection area, the denominator represents the union area. *IoU* and *c-IoU* only differ for plural phrases: *IoU* computes a union box (dashed) covering all components, while *c-IoU only considers* the area of the individual components to compute the intersection and union.

Plural phrases describe multiple entities in an image, either through a collective term (e.g. *crowd*) or a plural form (e.g. *two children*). Depending on the annotation, the gold box consists either of a single box enclosing all entities or several component boxes representing the individual entities. By convention[1], component boxes are merged into one *union box* spanning all individual boxes, functioning as a single gold box. Figure (2.a) gives an example of a union box for a plural phrase with two components. This reduction of multiple boxes to a single union box is widely established in PG evaluation, both for ground truths and predictions.

Although plural phrases are underrepresented in PG benchmarks, they constitute substantial proportions, and appropriate annotation and evaluation of component boxes is essential to achieve high-

[1]as, e.g., adopted in Plummer et al. (2015) and since then presumably adopted in the community for comparability

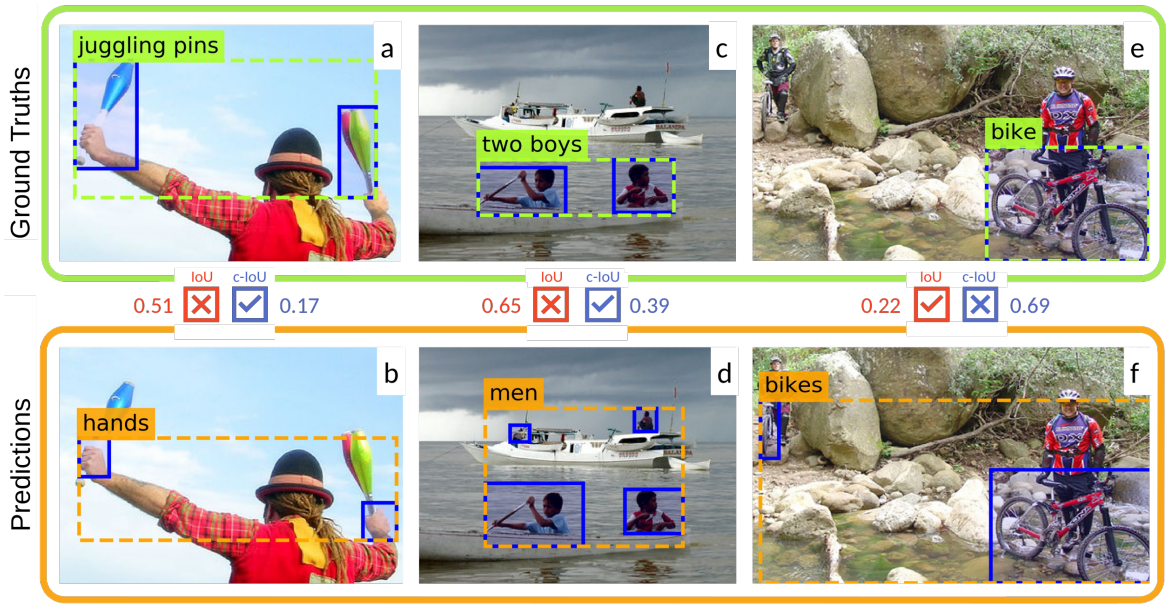

Figure 2: Ground truth (green/top) and prediction (orange/bottom) cases with components (blue) and union boxes (dashed). The Ground Truth box label in green represents the phrase being grounded; the orange phrase marked in the prediction represents the concept fitting the detected plural object. The scores represent the IoU score (red, left) and the c-IoU score (blue, right), respectively. Predictions with scores ≥ 0.5 are considered correct. The check boxes show whether the metrics correctly evaluate the prediction. For example, for sub-figures (a+b) the c-IoU score is 0.17 and so the prediction is considered incorrect (≤ 0.5), thus the c-IoU metric correctly evaluates that the detected *hands* constitute an incorrect prediction for *juggling pins* (blue tick).

quality mappings for all phrase types. However, the annotation of plural phrases is challenging, as shown in Testoni et al. (2020); Marín et al. (2020) who investigate how phrases can refer to groups of objects or several entities within a group.

(Semi-)supervised PG systems generally do not differentiate singular and plural phrases and always predict a single box (Li et al., 2019; Lu et al., 2020; Plummer et al., 2015). For multiple boxes with the same predicted label, either the largest box or the union box is returned. Thus, components are not individually evaluated, and the same metric, *Intersection over Union (IoU)*, can be uniformly applied to a prediction box for any phrase type. IoU computes the ratio of the *area of overlap* over the *area of union* between a predicted and a gold box and is usually thresholded at IoU ≥ 0.5.

While IoU is a simple and effective metric for evaluating 1:1 mappings, we claim that it is unsuitable for evaluating plural phrases. We show that the union box is in fact not an ideal gold representation for plural phrases: it can make the gold box overly large, especially when including areas that do not represent any components, and thus introduces an evaluation bias favoring large prediction boxes. Our contributions are as follows:

i) We detect, describe and *quantify an evaluation bias* in the grounding of plural phrases when applying standard practice of measuring *IoU* over *union boxes*, using an unsupervised PG system on the PG dataset Flickr30k.
ii) We propose a *novel evaluation metric* based on component boxes rather than union boxes.
iii) We show that the *new metric alleviates this bias* and reduces the evaluation failures.

2 Evaluation Bias

IoU is the standard evaluation metric used in PG and rewards predictions that highly overlap with their gold boxes. For a *plural phrase* that links to multiple ground truth boxes, a *union box* enclosing all components is generated, so the same evaluation metric can be used for singular and plural phrase types. However, we argue that this method introduces a considerable bias, which may result in unfair evaluations. When evaluating on union boxes, we ignore all information about the components' sizes and positions, and only consider the union box outline. If components are spread across the image, a union box can become much larger than the combined size of its component boxes, which makes them imprecise and ambiguous.

Figures (2.a) vs. (2.b) show an example of two-component union boxes that are highly overlapping – one for *pins*, the other for *hands*. Hence, a system that returns the prediction (2.b) for *juggling pins* will be unduly considered as correct. Similarly, for a prediction with too few or too many components, IoU often fails to detect such mistakes, as in (2.c) vs. (2.d). The ground truth (2.c) for *two boys* includes only two components, yet a system predicting four components (including the men on the boat) will still be correct according to IoU.

This type of 'false positive' arises from the generation of *union boxes*, in conjunction with the relatively forgiving nature of the IoU metric for large predicted boxes. Given such undesirable failure cases, we conjecture that IoU can lead to unwanted evaluation biases and we investigate whether it is a sound metric for evaluating plural phrases.

2.1 Quantifying the Bias

We verify and quantify the potential evaluation bias on GT annotations of Flickr30k and empirical system predictions. Depending on the distribution of component boxes across the image, the union box can be large, even if the components themselves are small. In Fig. (2.a) 75.47% of the union box area does not represent any component, so we term this area *filler space*. On the complete Flickr30k data, we compute an average of 3.6 components per plural phrase, which on average cover only 68% of the union box area, leaving one third (32%) of the space unfilled. For 24% of all union boxes, the filler space covers more than 50%, the gold box being twice as large as its components.

Hence, there is considerable potential for an evaluation bias to arise, as the IoU metric may unfairly favor large prediction boxes in two ways: i) overly large union boxes allow the prediction of wrong object types that happen to fall into the gold union box area; and ii) even if objects of the correct type are predicted, a large union box may be filled with too many or too few objects compared to the GT, and may still satisfy IoU $\geq$0.5. To verify this hypothesis, we perform experiments using an unsupervised PG system, capable of processing plural phrases.

2.2 Bias in Context of System Performance

Most existing PG systems are (semi-)supervised learners and need to be adapted to the special case of plural phrases: their object detectors need to deliver union boxes, instead of single-object boxes. Since plural phrases are much less frequent than singular phrases, this distributional bias may lead to poorer predictions for plural phrases. Recently, unsupervised PG systems have been proposed (Wang and Specia, 2019; Parcalabescu and Frank, 2020) that achieve competitive performance, but are not subject to such frequency biases. We thus perform our experiments with a system that replicates Wang and Specia (2019)'s approach.

The system[2] maps phrases to predicted bounding boxes using similarity rankings derived from word embeddings for the phrase and the candidate box labels. Since our object detector only detects single objects, we automatically generate plural objects that include several objects, by combining boxes with the same label.

We apply the system to a test set of 10k images with 5 captions each, containing 3.3 phrases on average. We ground ca. 141k phrases, including ca. 31k plural phrases (21.8%) and measure accuracy for the predicted bounding box(es) for a given phrase, using the IoU evaluation metric (with a threshold at 0.5) as success criterion.

Table (1.a) displays evaluation results for *all phrases* vs. *plural phrases* only, and in both cases we distinguish predicted boxes comprising *single* objects only vs. *all* objects (single and plural), for various settings: i) *upper bound* (row 1); ii) performance of our PG system in different settings (rows 2-4); and iii) manipulated predictions, i.e. *max box* and *random predictions* (rows 5-6).

i) Upper bound *Upper bound* represents the highest possible PG performance, computed as percentage of phrases with at least one detected object that matches the GT. Using only *single* objects, we find an upper bound of 72.34 for all phrases and 46.67 for plural phrases. The fact that single objects – which cannot constitute correct groundings for plural phrases – provide 'successful candidates' for nearly half the plural phrases, emphasizes that IoU is not a suitable metric for plural phrase grounding. When considering boxes with multiple objects as candidates, the upper bound increases by 2.82 percentage points (pp.) to 75.16 on all phrases and by 20.98 pp. to 67.65 on plural phrases, demonstrating that *plural objects* are an essential addition.

ii) PG system evaluation This setting also shows an increase when considering *plural objects*, with an increase for *all phrases* by 1.69 pp., and for *plural phrases* by 5.45 pp. Candidate pruning fur-

[2]Details of the system are given in the Appendix.

a) IoU metric	All Phrases		Plural Phrases	
	single	*all*	*single*	*all*
Upper bound [+prun.]	72.34	75.16	46.67	67.65
Unsupervised PG	47.94	49.63	31.15	36.60
– [+pruning]	-	53.36	-	56.46
– [+pruning, +enlarged]	47.99	52.03	33.84	52.45
Max box predictions	23.63	23.63	32.19	32.19
Random predictions	17.97	20.75	24.09	29.17

b) c-IoU metric	All Phrases		Plural Phrases	
	single	*all*	*single*	*all*
Upper bound [+prun.]	72.34	73.69	46.69	60.94
Unsupervised PG	48.05	49.94	31.00	37.37
– [+pruning]	-	52.38	-	50.09
– [+pruning, +enlarged]	47.26	51.02	29.84	46.95
Max box predictions	21.45	20.98	22.88	22.19
Random predictions	9.17	13.86	7.62	14.08

Table 1: PG performance in accuracy computed with **IoU** vs. **c-IoU** on *all* vs. *plural phrases*, considering *single* object boxes vs. *all* (single & multiple) object boxes. *+pruning* filters candidates: for plural phrases we consider plural objects only; for singular phrases only single objects. *+enlarged*: size of detected objs. increased by 50%.

ther increases accuracy by 3.73 pp. on *all phrases* and 19.86 pp. on *plural phrases*, while limiting the potential exploitation of large candidate boxes.

iii) Manipulated predictions We hypothesized that using IoU with union boxes is too forgiving and favors large predictions, so we conduct experiments where we generate overly large object predictions: in one, predictions cover the entire image (*max box predictions*); in the other, original predictions are enlarged by 50%. For the *max box predictions*, we obtain overall 23% correct predictions, and 32% for *plural phrases*. Thus, every third plural phrase benefits from very large prediction boxes. Ideally, IoU is designed to punish predictions that are overly large or not well placed over the gold box, due to division by union area. However, the image frame limits the maximum box size to the size of the image, which reduces the normalization effect for large objects.

Measuring PG performance with enlarged prediction boxes [+enlarged], increases accuracy by 2.69 pp. for *plural phrases* when considering singular objects only – despite singular objects being unsuitable predictions by definition. This further supports our hypothesis that large predictions are generally favored. However, larger predictions do not increase performance on mixed phrase types, so singular phrases must be less affected by this bias. For plural objects, PG performance even decreases with enlarged predictions, which suggests that plural objects cannot benefit from expansion.

In sum, the high upper bound with singular objects for plural phrases, the strong performance when predicting the entire image, and the effect on PG performance when enlarging prediction boxes all support our hypothesis that *the evaluation of plural phrases by IoU is biased*. Hence, a new metric is needed to counter this bias.

3 Our new Evaluation Metric c-IoU

We aim at a metric that is not based on the union box, but its components. However – any metric is only as good as the quality of its underlying ground truth. When studying the annotation of plural phrases in Flickr30k, we found that many of them are imprecise or incomplete. Nearly one third of plural phrases are annotated with a single bounding box without components and for 9% the number of components does not match the cardinality of the referring phrase (e.g. two component boxes for *three women*), leaving 37% of the plural phrases without proper representation of their components. This high level of noise precludes any metric that relies on matching the number of component boxes of ground truth and predictions.

In §2.1 we identified the filler space of union boxes – jointly with IoU evaluation – as the source of the detected evaluation bias. To combat this, we define an *adapted IoU* that is not computed over the union box, but its *aggregated components*, by taking the intersection of all gold and predicted component boxes and dividing by the area of the union of all (gold and predicted) component boxes. We call this metric *component IoU (**c-IoU**)*. c-IoU is analogous to standard IoU for single-object boxes, and only affects the evaluation of plural phrases, as seen in Fig. (1). By considering the area covered by *all* component boxes, it gains robustness against annotation noise.

Fig. (2.a-d) show two examples where IoU fails to correctly evaluate predictions, while c-IoU succeeds. The prediction *hands* for the phrase *juggling pins* yields an IoU score of 0.51, which accepts the prediction. The c-IoU score of 0.17, by contrast, rejects this prediction. Similarly, the prediction *men* is considered correct for *two boys* by IoU (0.65), but correctly rejected by c-IoU (0.39).

4 Evaluation of Component IoU (c-IoU)

We evaluate **c-IoU** in the same way as we did in §2.2 to quantify biases under IoU, and give results in Table (1.b). We expect c-IoU to avoid biases for plurals and ensuing false predictions.

The experiments confirm our expectation: large prediction boxes yield lower scores with c-IoU. *Max box predictions*, computed on all objects found in the image, yields 9.31 pp. lower accuracy on plural phrases. PG system performance with enlarged predictions measured on singular objects for plural phrases increases by 2.69 pp. to 33.84 for IoU, yet decreases by 1.16 pp. to 29.84 for c-IoU. Therefore, c-IoU better detects wrong predictions.

But c-IoU does not catch all incorrect predictions: with pruning, accuracy measured with c-IoU increases less (+12.72 pp. to 50.09) compared to IoU (+19.86 pp. to 56.46). Data inspection shows that without pruning, c-IoU allows plural objects for singular phrases (and vice versa) – typically the plural object includes the targeted object plus another small object in the background (see Fig. (2.e+f)). Since the background object drastically expands the union box but not the component union area, IoU may correctly evaluate such cases, while c-IoU could fail. Hence, a combination of both metrics could be beneficial, where c-IoU ensures that the right components are selected, while IoU may detect out-of-focus objects.

As a final test, we evaluate both metrics on artificially generated false plural box predictions. For each phrase, we assemble a *random prediction* consisting of 2-5 components with different labels. Ideally, most predictions should be labeled as incorrect, thus a lower accuracy indicates a more sensitive metric. c-IoU indeed returns much lower scores than IoU: 9.17 and 7.62 (c-IoU) vs. 19.97 and 24.09 (IoU) on *all phrases* and *plural phrases*, showing that c-IoU effectively counters the bias.

5 Conclusion

We have detected, described and quantified an evaluation bias for plural phrases in the PG literature. Our alternative c-IoU metric, acting on components rather than union boxes, alleviates this bias, as we show in experiments with an unsupervised PG system. Future work could test more systems to assess by how much state-of-the-art performance is lower than currently estimated. Evaluation of plural phrases is further impeded by the low quality of the gold boxes. Therefore, future benchmarks need to annotate plural phrases with all their components if we wish to enable PG systems to better learn the intricacies of language (including plural expressions) in relation to the visual modality.

References

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338.

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. Tell me more: A dataset of visual scene description sequences. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 152–157.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. 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.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual Genome: Connecting Language and Vision using Crowdsourced Dense Image Annotations. *International journal of computer vision*, 123(1):32–73.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. *arXiv preprint arXiv:1908.03557*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446.

Nicolás Marín, Gustavo Rivas-Gervilla, and Daniel Sánchez. 2020. A preliminary approach to referring to groups of objects in images. In *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–7. IEEE.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Letitia Parcalabescu and Anette Frank. 2020. Exploring Phrase Grounding Without Training: Contextualisation and Extension to Text-Based Image Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 962–963.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k Entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Alberto Testoni, Claudio Greco, Tobias Bianchi, Mauricio Mazuecos, Agata Marcante, Luciana Benotti, and Raffaella Bernardi. 2020. They are not all alike: Answering different spatial questions requires different grounding strategies. In *Proceedings of the Third International Workshop on Spatial Language Understanding*, pages 29–38.

Josiah Wang and Lucia Specia. 2019. Phrase Localization Without Paired Training Examples. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4663–4672.

A Appendix

System details Our approach replicates the unsupervised *bag-of-objects approach* by Wang and Specia (2019). The phrase and the labels for candidate objects are embedded using 300-dimensional word2vec embeddings (Mikolov et al., 2013). The object candidates are ranked by their *cosine similarity* to the phrase, and the object with the most similar label is returned. If there are several objects for the highest ranking label, we return the largest one in case of singular phrases and the union box (plural object) in case of plural phrases. In contrast to prior systems that used (multiple) object detectors with large label sets (545 or 1600 labels), our object detector, trained on Visual Genome (Krishna et al., 2017), uses only 150 coarse-grained labels.

We test performance on the Flickr30k Entities (Plummer et al., 2015) dataset for phrase grounding. The test set consists of 10k images with 5 captions each, containing 3.3 phrases on average. We ground 140 972 phrases, including 30 762 plural phrases (21.8%). The vocabulary of the phrases is relatively diverse with 8301 different words on the test split.

Evaluation examples Fig. (3) shows a few more examples of ground truths and our system's predictions, as well as the correctness of the evaluation using IoU and c-IoU. Fig. (3.a-d) show examples where IoU accepts predictions with incorrect object labels, while c-IoU rejects them. In (3.e+f), c-IoU finds that two hats are missing while IoU accepts the incomplete prediction. For example (3.g+h), c-IoU fails to identify that the prediction has a missing component. IoU correctly evaluates the prediction as incorrect because of the obvious union box difference, which makes the IoU drop below 0.5. Example (3.i+j) is a challenging case, as the phrase [*two young men clutch rags in]* ***their hands*** requires context for correct grounding, which is not provided by a PG system that looks at phrases individually. As expected, our system additionally predicts the old man's hand, which is incorrect but since the superfluous hand has a small area and is located closely to the others, both evaluation metrics fail to detect this mistake. Finally, in (3.k+l) the ground truth is missing the annotation of the components, so that c-IoU cannot correctly evaluate this correct prediction.

Figure 3: GT and prediction cases with union boxes (dashed) and components (blue); the check marks show whether IoU (red) and c-IoU (blue) correctly evaluate the prediction (for further explanation see Figure 2).