This CVPR workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

The Topology and Language of Relationships in the Visual Genome Dataset

David Abou Chacra
University of Waterloo
Ontario, Canada
d2abouchacra at uwaterloo.ca

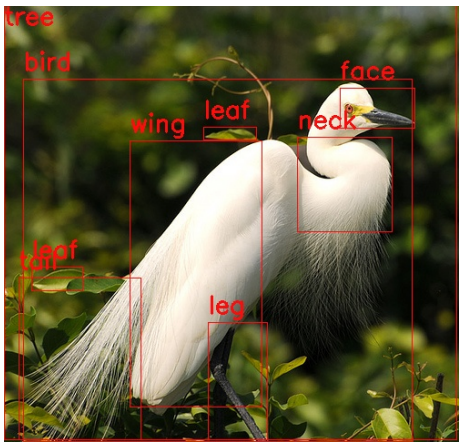
John Zelek
University of Waterloo
Ontario, Canada
jzelek at uwaterloo.ca

Abstract

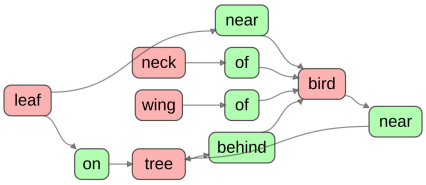
The Visual Genome Dataset is the de facto standard dataset used in Scene Graph generation. It contains a large collection of images with corresponding object and relationship labels. We explore the lingual aspect of the relationship predicates and find that very few symmetric/inverse relationships are represented in the dataset(for example, 'above' and 'under'). We believe this is linked to human spatial cognition, and posit that labelling bias stemming from human representations of relationships creates asymmetric relationship labels that span the whole dataset. We also perform a 2D topological analysis of the bounding boxes linked by different relationship predicates. This analysis sheds light on certain classes and their ambiguity wherein more frequent classes are semantically overloaded and therefore quite confusing. Finally we show that when reduced to more lingually and topologically well defined spatial relationships scene graph generation algorithm performance improves tremendously, but scene graph generators are still far from perfect.

1. Introduction

The Visual Genome (VG) dataset [7] is a collection of over 100,000 human annotated images that has been used extensively in computer vision research. A main motivation for creating the dataset was to allow for more cognitive-based computer vision research that is focused on image understanding and reasoning, rather than solely image perception tasks such as object detection or image segmentation. VG enables research that incorporates this sort of reasoning such as Scene Graph Generation [12, 13, 20], Visual Question Answering [14, 19], Image Captioning [5], among others [1, 16]. The full VG dataset is composed of a collection of 108K images, along with human generated annotations in the form of class-labelled bounding boxes around the objects in the images, attributes describing those objects, relationships between those objects, as well as question-answer pairs about the images. A sample of some of the kinds of data found in the VG dataset is shown in Figure 1.



(a) The sample image with labelled bounding boxes.



(b) The labelled relationships between the objects(objects displayed in red, while relationships are in green).

Figure 1. A sample of a data point in the VG dataset, the object relations are displayed in the form of a scene graph. Object attributes are not shown for clarity.

1.1. The VG dataset and Scene Graphs

The images comprising the VG dataset were taken from the YFCC100M [15] and COCO datasets [9] and then annotated rigorously using human annotators crowdsourced though an online platform. In short, labellers were tasked with creating text descriptions of regions in the image, these text descriptions are then grounded into the specific parts they're describing using bounding boxes to ground the objects being described and relationships and attributes being connected to and between the object bounding boxes. The

final dataset is comprised of over 3.8 million total bounding boxes classified into 33,877 object categories, these bounding boxes are connected by over 2 million total relationships (classified into 42K distinct relationship predicates), in addition to over 2.5 million attributes describing the classified objects in the bounding boxes (with 68K distinct attributes). On average, one image is expected to contain 35 object bounding boxes, 26 attributes and 21 relationships.

The Visual Genome Dataset therefore lends itself very well to the task of scene graph generation [3, 12, 13, 20], where given an input image, a model is expected to output the objects found in the image as well as describe the relationships between them. In this task, the vast amount of objects and relationships found in the VG dataset can be a drawback due to severe class imbalance across object categories and relationship predicates. It is common practice across scene graph literature to instead opt for using a subset of the VG dataset, the VG200 dataset [12], containing the 150 most frequently occurring objects along with their 50 most frequent relationships. The final object count in the VG200 dataset is 1,145,398 objects, i.e. the top 150 object classes (out of the 33K classes) accounted for approximately a third of the total bounding boxes. The total preserved relationships in the VG200, which are spread across 50 predicates is 622,705 relationships (these are out of the original 42K predicates that described the 2M original relationships). Overall, this serves to lessen the severity of the inherent class imbalance across objects and relationships, though not entirely, without altering the original Visual Genome Dataset too severely. Another common practice in scene graph literature is formalizing relationship triplets found in the VG dataset as [subject, predicate, object] triplets. For example, one such triplet observed in Figure 1 is [leaf, on , tree], where leaf is the subject, on is the predicate and tree is the object of the relationship.

2. Language and Inverse Relationships

In language, spatial prepositions often have inverses, which can serve as a dual(but opposite) representation of the same physical phenomenon being observed. For example, if a table is 'to the right of' a person, it is immediately understood that the person is 'to the left of' the table. Several of the 50 predicate classes that exist in the VG200 data set have linguistically 'inverse' relationship predicates within the set as well. For example, the predicates 'behind' and 'in front of' are both in the VG200 predicate set. It follows that if a subject-predicate-object triplet of subject A-'behind'-object B exists, we would expect to see the inverse triplet of subject B-'in front of'-object A. One very commonly occurring predicate, 'near', could even function as its own inverse.

Figure 2 shows a heat map of how often two relationship predicates share an inverse relationship in the VG dataset.

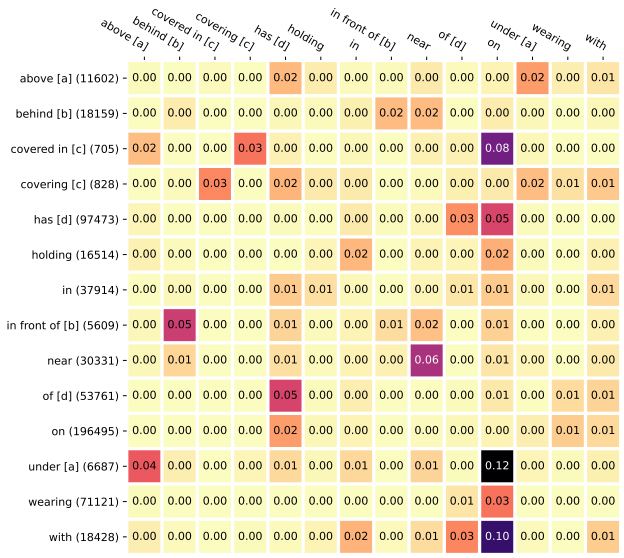


Figure 2. A heatmap of the occurrence of inverse relationships for specific predicates. The letters in square brackets indicate which predicates we expect to be inverse pairs, the numbers in parentheses are the total occurrences for this predicate in the dataset. For every row the value in the heat map reflects the ratio of: (inverse relationship occurrences of the row predicate with the predicate in the column) to (total occurrences of the predicate in the row).

An inverse relationship exists if the same two bounding boxes(containing the same specific objects) share two relationships, with one in each direction. In other words, one inverse relationship exists between predicates 'above' and 'under' if for a specific pair of objects the objects are linked by the triplet subject A-under-object B as well as subject B-above-object A in the dataset. Note that two objects may have multiple relationships connecting them. The full heatmap of inverse relationships between all 50 predicates can be found in the supplemental material.

We notice that inverse relationships do not form a significant portion of the relationships observed in the VG dataset. In fact, even the expected inverse relationships between linguistically inverse predicates are not at all frequent. Predicates 'under' and 'above' (or 'under' and 'over' whose result can be seen in the supplemental material) don't share much of an encoded inverse relationship, in fact 'under' seems to share a stronger inverse relationship with 'on', however that is likely also due to how over-represented the predicate 'on' is in the VG dataset.

The work done by Landau and Jackendoff [8] on human spatial cognition touches on a relevant issue. They describe the 'asymmetry' in the way humans form spatial representations, where these asymmetries come from many factors, including that certain objects are more likely to be the 'reference point' based on size or relevance or saliency.

Even the more apparently 'symmetrical' **spatial** predicate relations tend to become asymmetric in our reasoning by virtue of how humans form their own spatial reasoning. Given that the VG labels are generated by human annotators, there is an asymmetric skew that will inevitably exist in the resulting labels which is the root cause for why these inverse relationships do not exist. For example, of the over 243,000 relationship triplets that include humans in the VG200 dataset, humans are the subjects in approximately 84% of those relationships, while they are objects in only 19%.

This asymmetric skew is to be expected in human conversation and description, mainly because human reasoning can understand the inverse relationship immediately and it does not need to be explicitly stated. However, this will not be the case for learning algorithms who don't have an existing knowledge of the world, or of the verbal semantics of the relationship predicates they are predicting. A small percentage of inverse relationships existing in the VG200 dataset, even for the predicate classes where we expect them to exist, could likely hinder the ability of learning models to understand these relationships. A potential solution to this could be in the form of data augmentation (for a data-driven solution) or even prior knowledge about these inverse relationships being given to the learning algorithms utilizing this data. Alternatively, inverse relationships could be used as a metric for measuring generalization performance of learned models, especially if certain models were shown to be able to piece together these inverse relationships without explicitly being told about them, or incentivised to learn them.

It is worth noting we also measured co-occurring 'forward' relationships between the same two objects i.e. two objects related in the same subject-object configuration but with different predicates. This measurement yielded no results of interest, as these relationships turned out to be very rare.

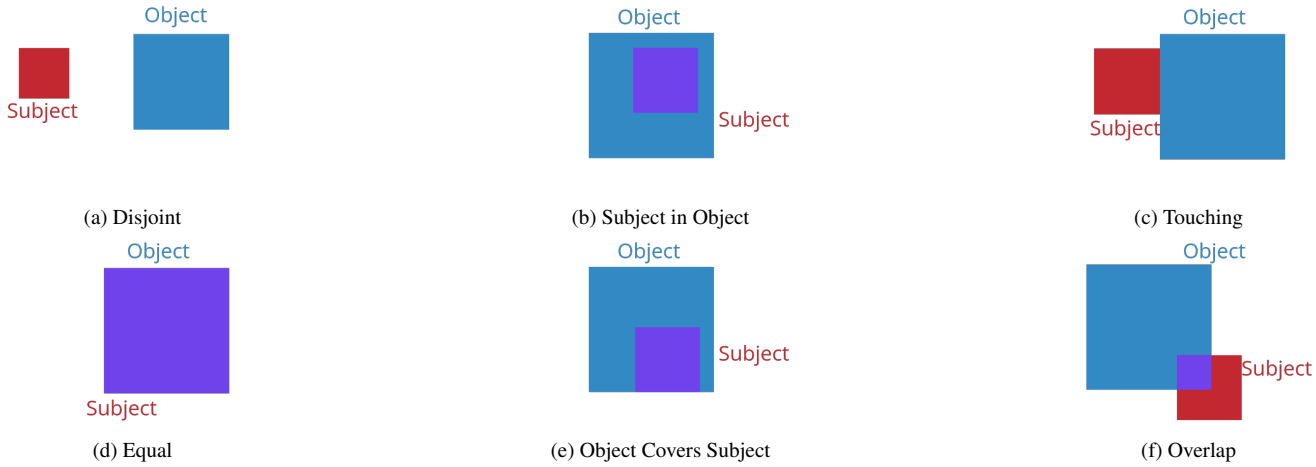
3. Topological Relationships

The language that creates the relationship triplets may be biased by how humans view and reason about the world, which makes the bounding boxes that also define these triplets worth exploring as well. These bounding boxes are the smallest 2D image axis-aligned rectangles that can hold the object they border and they lend themselves well to a 2D topological analysis. Topological relationships [2, 4] can be determined between two 2D areas, and the topological relationship can be classified depending on the configuration between the borders and the interiors of these areas. Figure 3 describes the possible topological relationships between two 2D areas.

A topological analysis of the bounding boxes found in the VG dataset sheds light on the relationships in the dataset. Where our language and how we describe a relationship can be influenced by our cognitive biases, observing the topological relationships between the bounding boxes can give us an understanding of what a certain relationship is prioritizing. They can inform us on whether the subject or the object is the more 'dominant' for a given predicate class as well as validate whether the downstream task of scene graph prediction that utilizes the features in these bounding boxes is being built on valid data. Since several scene graph generation approaches [1, 18] operate by taking the union or intersection of the detected object bounding boxes to predict the relationship predicate, a topological perspective on how these bounding boxes are related in the VG dataset is quite relevant.

Furthermore, we analyse the dominant directions in which these relationships are occurring. These directions are found by analysing the location of the object relative to the subject when they are linked by a specific relationship predicate. For triplets with predicates describing spatial prepositions, such as [subject, 'above', object], we expect to see the object always being towards the south of the subject. This analysis also sheds light on whether more frequent and more vague predicates (such as 'on' or 'has') are exhibiting any regular directional relationships.

We visualize some of the results of the topological analysis in Figure 4 and the directional analysis in Figure 5. The full heat maps for all the relationship predicates can be found in our supplemental material. Note that while the 'equal' topological relationship doesn't occur in this subset of predicates, mainly due to its more specific and rare configuration, it is does show up in the full set. Also noteworthy is our evaluation of the 'covers' versus 'in' topological relationships. While [4] describes these relationships rigidly(as shown in Figure 3), we loosened them very slightly (in the order of 5% of the smaller of the two bounding boxes under analysis) to account for human error in labelling the bounding boxes. For example, if a subject A lies completely within object B for a given case, however it is proximal enough to the boundary of object B (though not exactly touching it as shown in Figure 3e) we could still bin the topological relationship as a 'subject covers object' relationship depending on how close subject A is to the boundary (of B) relative to its own size. Our directional calculations are binned into the 8 cardinal directions of a compass, and measured based on the relative centers of gravity of the bounding boxes. For example if for a certain relationship triplet **[subject, predicate, object]** the center of gravity of the **object** bounding box is **south-east** of the center of gravity of the **subject** bounding box, it is binned as 'SE'. We also calculate the spread of the directions for the 8 different topological configurations within each predicate, and isolate some predicate-topological pairs of interest where directions exhibit a noteworthy spread, this is shown in Figure



(a) Disjoint (b) Subject in Object (c) Touching

(d) Equal (e) Object Covers Subject (f) Overlap

Figure 3. Topological relationships [2, 4] visualized. Note that two additional relationships exist Object in Subject (similar to (b), but with object and subject reversed), as well as Subject Covers Object (similar to (e) but with the object and subject reversed).

6.

The topological relationships tend to reveal the more 'dominant' of the subject and object pair linked by a certain predicate. In some cases, such as the predicate 'has'(e.g. building has window), the expected topological configuration is dominant: the object, window, is fully contained in the subject. This follows from how we expect the lingual relationship to occur [8]. Notably, the topological spread of the predicate 'in' is not as would be expected, and further highlights the vagueness of this relationship predicate. While the expected dominant topological configuration (subject in object - it is literally in the name) is the most frequently occurring, it is **not** extremely dominant. We would expect a 'subject in object' topological configuration for the relationship triplet [person in car], but, for example, the triplet [bottle in hand] (where the bounding box of the bottle is actually larger than that of the hand produces an 'object in subject' topological configuration, and an example triplet [plant in pot] counter-intuitively produced a disjoint topological configuration due to how the bounding boxes are labelled.

The directional evaluation produced more expected results. Predicates 'above' and 'under' mostly exhibited directional configurations that are true to their descriptions. In fact, their symmetric relationship is highlighted well by how they exhibit similar topological configuration distributions, while having inverse directional configurations. A similar, but less pronounced, symmetric relationship is also seen in the predicates 'behind' and 'in front of'. Vague predicate classes, however, such as 'on' or 'in' still showed a big variety of directional configurations, likely due to them encoding several different lingual interpretations of 'on' and 'in'. The predicate 'near' interestingly seemed to imply the subject and object were side by side (with the slightly higher chances for a 'W' and 'E' configuration). The results shown in Figure 6 help shed some light on the combination of topological and directional configurations and serve to disambiguate some predicate classes. For instance, the predicate 'on' exhibited more predictable directional qualities when the topological relationship was 'overlap'. In this predicate + topological combination, 'on' usually meant the subject was on top of the object e.g.[person on sidewalk], as opposed to 'on' with the configuration 'Subject in Object'(e.g. [fruit on tree]) where the subject is potentially anywhere within the bounding box of the object.

Similarly to how inverse relationships can be used to augment the dataset, we believe it is possible to modify the more vague relationship classes based on their topological configurations. Spatial predicates that are linguistically similar and exhibit similar topological and directional configurations could possibly be merged into broader classes without losing too much of their meaning. For example, predicates such as 'laying on', 'lying on', 'parked on' which all occur in the VG200 dataset, and all seem to be describing a similar spatial configuration (further proven by their topological configurations) can be merged into a super set[1]. While on the other hand the larger and vaguer predicate classes 'on' or 'in' can possibly be broken down.

We would also like to note that a topological analysis of the bounding boxes in VG may be subject to certain biases and shortcomings as well. We live in a 3D world, and it may be difficult for any computer vision system to infer the 3D concepts from 2D images in the Visual Genome dataset annotated with 2D bounding boxes. Concepts like

[1]This could provide an alternative to the synset embeddings that are extracted from Wordnet [10] IDs which are already supplied in the VG dataset. Note we did not evaluate the topological configurations while utilizing those IDs.

Figure 4. A heatmap of the occurrence of topological relationships between bounding boxes related by specific predicates. The values shown in the heatmap are the portions of the total occurrences of the row predicate that exhibit the specific topological configuration in the column. The values in parenthesis next to the predicate names are the total occurrences of that predicate.

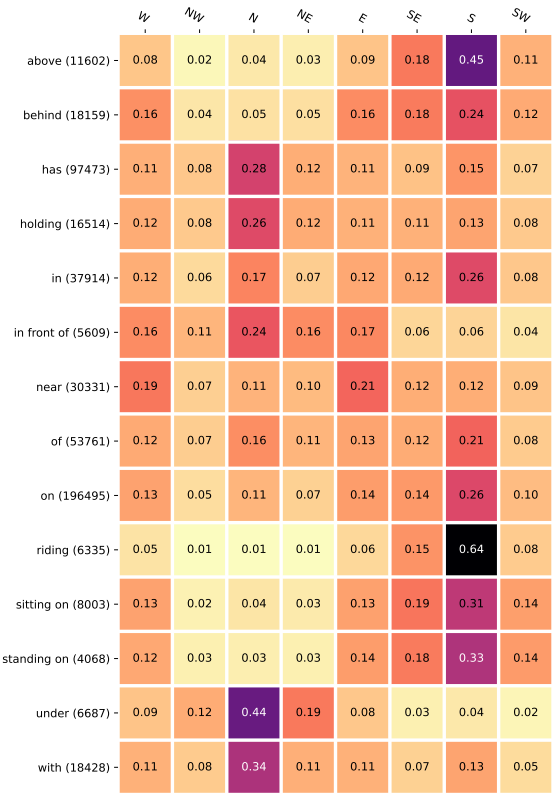


Figure 5. A heatmap of the angles between subject and object for selected relationship predicates. The values shown in the heatmap are the portions of the total occurrences of the row predicate that exhibit the specific directional configuration in the column. The values in parenthesis next to the predicate names are the total occurrences of that predicate.

'behind' and 'in front of' may be extremely difficult for a vision system that has only seen 2D images to reason about, especially if it is not designed with the 3D world in mind. A topological analysis of the VG dataset is likely better suited for the relationship labels that are not overtly 3D in nature. Relationships like 'above' or 'under' are more two dimensional than 'in front of' or 'behind', for example, which may be why the symmetric relationship between the more 2D pair(above-under) was more easily distinguishable in the topological and directional analysis than that of of the more 3D pair(behind-in front of). With that in mind, we still see value in this analysis and the properties that it was able to reveal in the underlying data.

4. Algorithmic Use in Scene Graphs

While we believe the topological and directional configurations along with the augmentations defined by spatial language discussed previously could possibly be incorporated into a novel algorithm for scene graph generation, it is outside of our scope of discussion in this work. Instead we aim to experiment with different data configurations based on what our exploration has yielded. The topological and lingual analysis enabled us to better understand the ambiguities of the labels and restructure relationships in a manner that is lingually and topologically sound. We created 2 alternate subsets of the VG200 relationship predicates and measured the performance of the same baseline model when trained with these new labels.

In this section we conclude with 3 simple scene graph generation experiments that are driven by modifying the data rather than modifying the underlying algorithm. To reiterate, in scene graph generation [1], we are given an input image and tasked with identifying the objects in that image along with the relationships that exist between those objects much like the graph shown in Figure 1. Scene graph generators are evaluated under 3 different settings:

- **Predicate Classification**: Where the object bounding boxes and the object class labels associated with the

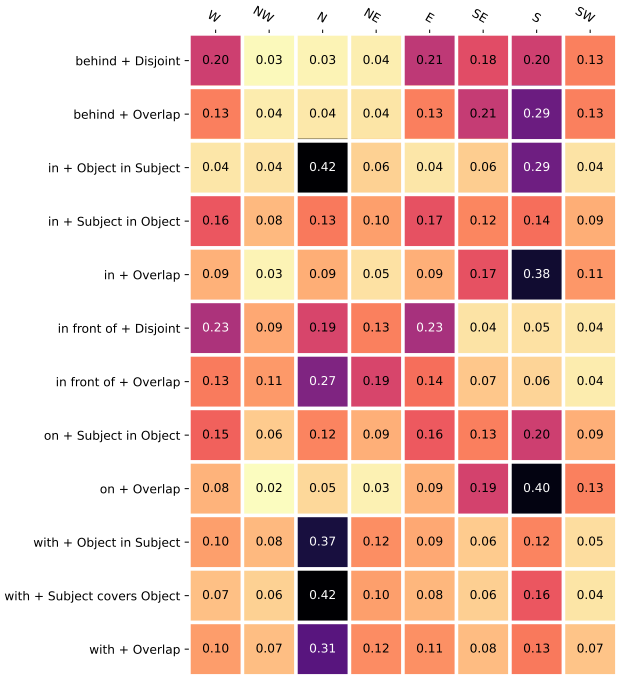


Figure 6. A heatmap of the angles between subject and object for selected relationships and specific bounding box topologies for bounding boxes that are connected by those relationships. The values shown in the heatmap are the portions of the total occurrences of the row predicate+topological configuration that exhibit the specific directional configuration in the column.

bounding boxes are given, hence the scene graph generator must only find the **relationships** between the given objects.

- **Scene Graph Classification**: Where the object bounding boxes are given, but the class labels associated with the bounding boxes are not, so the scene graph generator must infer both the **class labels** as well as the **relationships** between the bounding boxes.
- **Scene Graph Generation**: Where the input image is given without any other labels or information, and the scene graph generator must uncover the relevant **objects** in the image, their **bounding boxes** as well as the **relationships** between them. This is the most challenging setting for evaluation.

Scene graphs are also evaluated based on their recall, as opposed to based on their precision. Given an input image, the scene graph generator is evaluated on how many of the ground truth relationships it was able to uncover. The defining metric in scene graph literature is the mean recall@K metric. The mean recall averages the recall score across every predicate class individually instead of every predicted relationship instance. This is mainly due to the large predicate class imbalance that exists in the VG data set. So the average recall is calculated for every predicate separately first, and then averaged again to get the mean recall which ensures under-represented predicate classes are not being ignored in the evaluation. The mean recall@K metric is the mean recall score when the top K scene graph predictions are used for evaluation, so a mean Recall@20 would mean the scene graph was allowed to predict up to 20 relationship triplets to compare to the ground truth.

The baseline scene graph generator we utilize for the experiments is the Stacked Motif Network (MOTIF) [20]. In brief, the Stacked Motif Network (MOTIF) [20], first generates the object label only then utilizes a bidirectional LSTM to propagate information between the different object proposal and relationship proposal stages, effectively allowing object context to influence its label and its relationship labels. For our experiments we follow the implementation of [6] and exchange the VGG16 [11] detector with a RESNeXt-101-FPN [17] which was shown to improve performance. Proposing a novel scene graph generation model is out of the scope of this work and we only aim to see the differences in performance that a strong baseline generator can observe when the data it uses is better structured. Scene graph generation networks usually achieve relatively low recalls (with the Scene Graph Generation mean Recall still being under 10% in state of the art models [1]). The reasoning authors give is usually the vagueness of the predicates and 'long tailed'-ness of the distributions in the VG dataset. Certain relationships dominate the dataset and learning algorithms struggle to capture the true conceptual information contained in the entirety of the dataset, instead focusing on the dominant classes. As shown in our lingual and topological analysis, the VG dataset does indeed show topological and lingual ambiguity, lack of symmetrical relationships and labelling bias, and these are detrimental to learning models.

We train and evaluate the same scene graph model [6,20] on three different predicate configurations derived from the VG200 set:

- **Original Predicates**: We use the original 150 classes and 50 predicates from the VG200 data set to baseline the model.
- **Relationship Subset 1**: The 'less vague' subset where we remove 14 of the more vague original 50 predicate classes, and merge 4 others to keep 32 unique relationship predicates. The removed classes are either linguistically vague, or did not exhibit topological and directional configurations that matched their descriptions. That being said, we do keep the larger vague classes (such as 'in' or 'on') since they form such a large subset of the dataset.
- **Relationship Subset 2**: The spatial preposition sub-

set where we take a subset of the 50 classes that correspond only to spatial prepositions, we also merge classes that exhibit similar lingual, topological and directional configurations. We end up with 8 unique predicates that are a combination of 20 of the original predicates.

The exact predicates we use are listed in Table 1. The results of the experiments are shown in in Table 2. At first glance the recall results when using relationship subset 2 may seem to indicate a significant leap in performance, though we also understand that this leap is very much expected as the class labels are better balanced and much fewer. That being said, it's interesting to see that even an off the shelf scene graph generator can perform quite well as a spatial preposition predictor when given the right data. In our opinion the more interesting result is that of relationship subset 1. This experiment showed some improvement in recall with the reorganized 50 relationships, but that improvement is not as significant as we would have expected. Relationship subset 1 cleaned up edge cases and some of the more 'vague' predicates of the original 50, but the performance improvement seen was relatively small.

We analyze the existence of inverse relationships in the predictions of the scene graph generators and show a subset of the results in Figure 7. In this analysis, we tallied what inverse relationships are found by the scene graph generator for each of the correctly recalled ground truth relationships. In other words, if a ground truth relationship is correctly found by the generator in its top K relationships under a specific setting, we find whether an inverse relationship is also being predicted(whether it exists in the ground truth or not). This yielded some interesting findings on what inverse relationships the generator is learning. For example, in the case of the original 50 predicate classes, the predictor seemed to find a strong inverse relationship between predicates 'of' and 'has', as well as 'on' and 'has'. Both of these pairs are likely a result of a symmetric possessive relationship that is getting encoded (e.g. wing of bird/bird has wing or car has wheel/wheel on car). This is likely due to the formulation of stacked motif networks [20] which honed on certain repeated 'subgraphs' in the ground truth. In the case of relationship subset 2 of spatial predicates, some inverse relationships are more prevalent(such as the large class of 'on' having an 'under' inverse relationship 15% of the time), however other incorrect relationships also show up (such as 'behind' being its own inverse).

If anything the results of both of our experiments seem to indicate that there is still much to be done in the field of scene graph generation even outside of the dataset domain. Stacked motif networks [20] are an impressive approach to generating scene graphs, that managed to push the field forward by paying attention to the underlying data. Since then, a few other works have taken interesting approaches as well.

Original Relationships	Relationship Subset 1	Relationship Subset 2
above	above	above
across	across	-
against	against	-
along	along	-
and	-	-
at	-	-
attached to	attached to	-
behind	behind	behind
belonging to	-	-
between	between	-
carrying	carrying	-
covered in	covered in	-
covering	covering	-
eating	-	-
flying in	*in*	*inside*
for	-	-
from	from	-
growing on	growing on	*on*
hanging from	hanging from	*on*
has	has	-
holding	holding	-
in	in	inside
in front of	in front of	in front of
laying on	laying on	*on top of*
looking at	-	-
lying on	*laying on*	*on top of*
made of	-	-
mounted on	mounted on	*on*
near	-	-
of	-	-
on	on	on
on back of	on back of	*on top of*
over	over	*above*
painted on	painted on	-
parked on	parked on	*on top of*
part of	-	-
playing	*using*	-
riding	riding	*on top of*
says	-	-
sitting on	sitting on	*on top of*
standing on	standing on	*on top of*
to	-	-
under	under	under
using	using	-
walking in	walking in	*inside*
walking on	walking on	*on top of*
watching	-	-
wearing	wearing	-
wears	*wearing*	-
with	-	-

Table 1. A breakdown of the relationships used in each of our 3 experiments. A '-' means the relationship from the original set was removed entirely. Italicised text is used to indicate that a relationship has been kept but its label was modified.

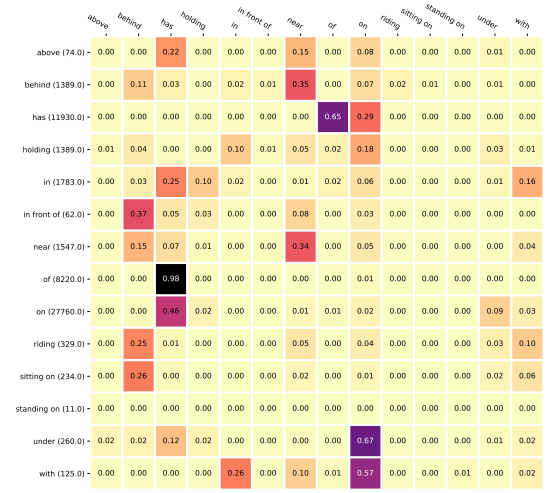
For instance, segmentation grounded scene graph generation [6] demonstrates the utility of moving away from solely bounding boxes for scene graph generation and shows that even mask annotations obtained via zero-shot transfer can improve scene graph generation performance. Grounding consistency [3] on the other hand demonstrated how a lack of negative training examples in the data and the reliance on recall alone in the evaluation led most scene graph predictors to learn very biased representations.

5. Conclusions

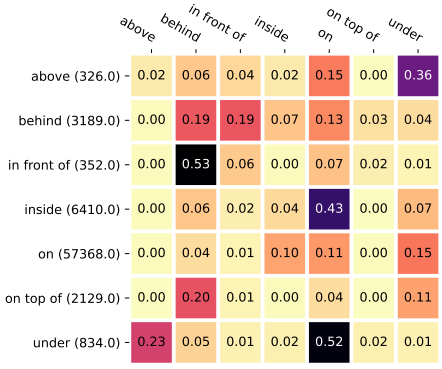
In this work we explored the lack of representation of both sides of symmetric relationships in the VG dataset, which likely resulted from the asymmetric spatial representations humans (and thus human labellers) exhibit. This means that inverse relationships such as 'above' and 'un-

Model	Detector	Relationship Set	Predicate Classification			Scene Graph Classification			Scene Graph Generation		
			mR@20	mR@50	mR@100	mR@20	mR@50	mR@100	mR@20	mR@50	mR@100
VCTree [13]	VGG-16 [11]	Original 50 Relationships (Reported in [6])	14.0	17.9	19.4	8.2	10.1	10.8	5.2	6.9	8.0
MOTIF [20]	RESNeXt-101-FPN [17]	Original 50 Relationships (Reported in [6])	14.1	18.0	19.4	8.0	9.9	10.6	5.8	7.7	9.0
MOTIF [20]	RESNeXt-101-FPN [17]	Baseline (Original 50 Relationships)	12.0	15.4	16.7	5.9	7.2	8.9	4.8	6.1	7.2
MOTIF [20]	RESNeXt-101-FPN [17]	Relationship Subset 1 (Less Vague 36 of Original Relationships)	17.4	21.2	22.6	8.5	10.2	10.7	5.8	7.7	8.9
MOTIF [20]	RESNeXt-101-FPN [17]	Relationship Subset 2 (Spatial Prepositions 20 of Original Relationships)	46.1	55.7	59.1	22.2	25.9	27.1	14.4	18.8	21.6

Table 2. The results of our three experiments with different VG200 predicate subsets. Our code implementation was adapted by starting with the implementations of [6, 12].



(a) Selected inverse relationships from the original predicate model.



(b) Inverse relationships from the model trained with only spatial prepositions (relationship subset 2).

Figure 7. Inverse relationship proportions in the scene graph predictors trained on different data subsets evaluated on the predicate classification task using the top 50 predictions. The numbers in parenthesis indicate the number of correctly recalled instances of each predicate, the numbers in the grid are the portion of those recalled instances that had an inverse relationship with the column predicate class (whether that relationship was in the ground truth or not).

der' are not encoded in the dataset, and could potentially be used as a measure of generalizability of models or to improve performance. We also discussed the topological and directional configurations exhibited by relationships in the Visual Genome Dataset and showed how overloaded predicate classes(such as 'on') exhibit topological vagueness which may confuse trained models. An interesting avenue for future exploration is a compositional analysis of the objects and relationships in the VG dataset. It would be quite interesting to see how well hierarchical relationships hold and whether more conceptual understandings can be garnered by utilizing a combination of language, topology and a heirarchy of parts.

References

[1] Xiaojun Chang, Pengzhen Ren, Pengfei Xu, Zhihui Li, Xiaojiang Chen, and Alexander G Hauptmann. A comprehensive survey of scene graphs: Generation and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1, 3, 5, 6

[2] Eliseo Clementini, Paolino Di Felice, and Peter van Oosterom. A small set of formal topological relationships suitable for end-user interaction. In *International Symposium on Spatial Databases*, pages 277–295. Springer, 1993. 3, 4

[3] Markos Diomataris, Nikolaos Gkanatsios, Vassilis Pitsikalis, and Petros Maragos. Grounding consistency: Distilling spatial common sense for precise visual relationship detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15911–15920, 2021. 2, 7

[4] Max J Egenhofer and Robert D Franzosa. Point-set topological spatial relations. *International Journal of Geographical Information System*, 5(2):161–174, 1991. 3, 4

[5] Jiuxiang Gu, Shafiq Joty, Jianfei Cai, Handong Zhao, Xu Yang, and Gang Wang. Unpaired image captioning via scene graph alignments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10323–10332, 2019. 1

[6] Siddhesh Khandelwal, Mohammed Suhail, and Leonid Sigal. Segmentation-grounded scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15879–15889, 2021. 6, 7, 8

[7] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalan-

tidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 1

[8] Barbara Landau and Ray Jackendoff. Whence and whither in spatial language and spatial cognition? *Behavioral and brain sciences*, 16(2):255–265, 1993. 2, 4

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[10] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 4

[11] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6, 8

[12] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3716–3725, 2020. 1, 2, 8

[13] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6619–6628, 2019. 1, 2, 8

[14] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6619–6628, 2019. 1

[15] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 1

[16] Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. Phrasecut: Language-based image segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10216–10225, 2020. 1

[17] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 6, 8

[18] Pengfei Xu, Xiaojun Chang, Ling Guo, Po-Yao Huang, Xiaojiang Chen, and Alexander G Hauptmann. A survey of scene graph: Generation and application. 2020. 3

[19] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731, 2019. 1

[20] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5831–5840, 2018. 1, 2, 6, 7, 8