TranstextNet: Transducing Text for Recognizing Unseen Visual Relationships – Appendix

1. Subject-Object to Relationship Mapping with a Text Generation Model

To generate text, we utilized the well-known GPT-2 [3] NLG model. We fine-tuned the model as described in the main text. We generated 23M sentences, which resulted in 19M triplets. To cleanup the generated text, from irrelevant objects, and parsing failures we projected the objects classes that were not in VG's object class set, by utilizing [2] word vectors and cosine distance. After cleanup, the Object-Relationship mapping `ORM` based on generated text included 12M triplets.

2. Subject-Object to Relationship Mapping with Novels from Project Gutenberg

We used text from Project Gutenberg [1]. We parsed 96M lines, which resulted in 82M triplets. We utilized the same cleanup procedure as described in the text generation section. After cleanup the books-based `ORM` included 26M triplets.

`ORM`	# Rows Parsed	# Triplets	# Relationships	# Object pair coverage
Image Based	6M	5M	1124	59%
Text Generation	23M	12M	10,953	62%
Book Based	96M	22M	18,984	69%

Table 1: A comparison of `ORM` based on different text sources.

3. Subject-Object to Relationship Mapping Comparison

A key point of this paper is the use of auxiliary text. As described in the main paper, we utilized three different sources of text. Refer to Table 1 (in this appendix) where we compare the three text sources: (1) sentences from image captioning datasets, (2) Gutenberg ebooks and (3) sentences generated through a natural language generation (NLG) model. The first column specifies how many rows were parsed, the second column shows how many relevant triplets were used after filtering, the third column shows the number of relationships in the `ORM`, and the last column is the percentage of all object pair combinations (22K) that have one or more relationships. Refer to Figure 3 for the top-19 relationships distribution of all the `ORM`s versus the relationships in VG-200. The diagram can provide an explanation to why leveraging auxiliary text helps reduce the training bias.

4. Additional Results

In Figure 1, (in this appendix) we present additional SGG results. All images present examples of seen relationships (in the yellow rectangles), and unseen relationships (in the blue rectangles. All images are results by TranstextNet$_A$ with Motifs [7] as an SGG backbone.

5. Elaboration on mR@K Metric

In the paper we demonstrated how ingesting text to SGG models improves recognition of less frequent relationships, while not compromising the detection of frequent relationships. Consider Figure 2 (in this appendix), where we present R@100 results for Motifs [7], with TranstextNet$_A$, and TranstextNet$_F$. The results clearly show that ingestion of text facilitates recognition of less frequent relationships.

6. Elaboration on Recognition of Unseen Visual Relationships

In the main text we reported the results of a new task called recognition of unseen relationships(ROUVR). This tasks the ability of SGG models to transduce relationships from the parsed text. We tested our model on the VRDD relationship set. The relationship set included 57 relationships unseen during training and 13 that were seen during training. During testing TranstextNet was able to recognize 42 unseen relationships from the data set. In Figure 1 the middle image for example: between the pair `person-1` and `kite-1`, two relationships were recognized, `touch`, and `holding`, which are semantically similar. This is a recurring theme

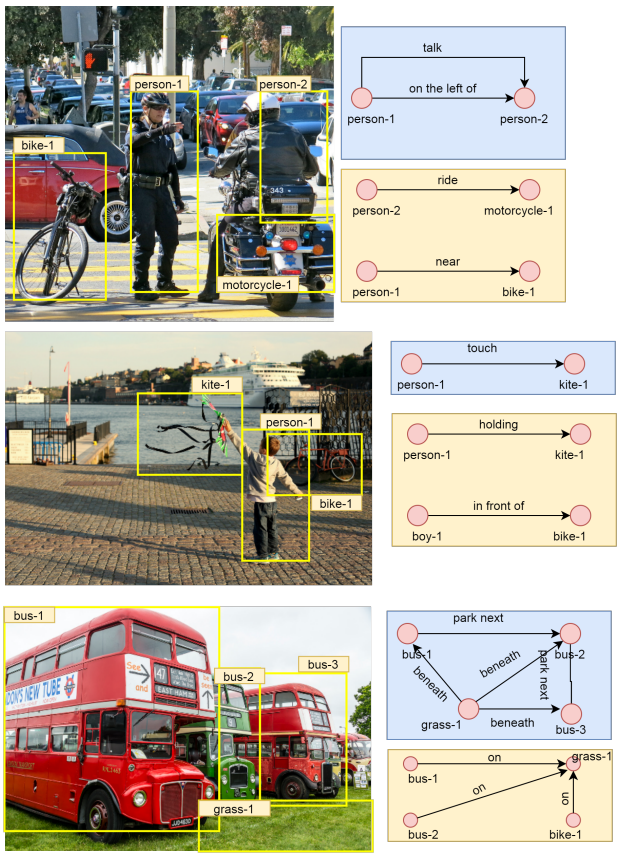

Figure 1: Qualitative results from TranstextNet with [7] as a backbone. The blue rectangles are unseen relationships recognized by TranstextNet and the yellow rectangles are of seen relationships. Images are taken from VRDD.

as our system has shown the ability to recognize synonyms of seen relationships, another example is in the bottom image, the relationship `beneath` is a synonym of `under`, which was seen during training. Refer to Figure 4 for a histogram of unseen relationships detected by TranstextNet on the VRDD test set. This histogram supports that TranstextNet successful in recognizing unseen relationships. The text generation model was fine-tuned with HugginFace transformers library [5]. We used 6M image captions to fine-tune the GPT-2 model [3].

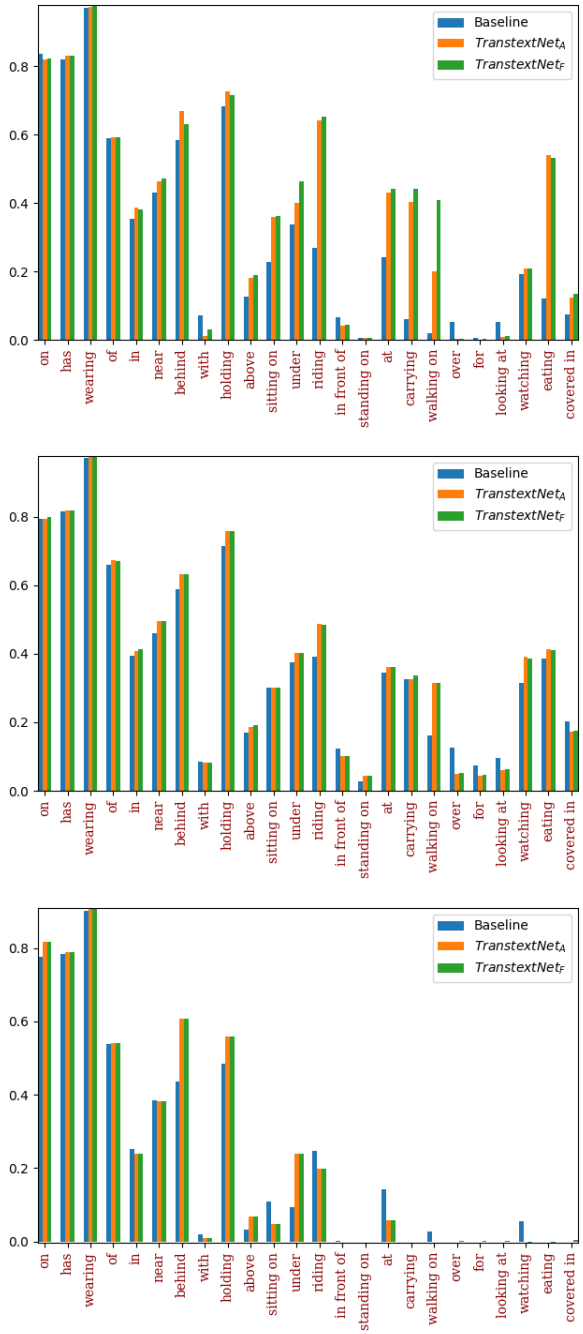

Figure 2: Comparison of all baseline models [7, 6, 4] performance on the top-25 relationships in VG. The metric used is R@100. The setup is Pred-Cls. In blue is the baseline model, in orange is TranstextNet$_A$, and in green TranstextNet$_F$. The left diagram shows the results of TranstextNet with Motifs model[7] as an SGG backbone, the right diagram shows the results with the VCTree model [4] as a backbone, and the bottom diagram shows the results with the IMP model [6] as a backbone.

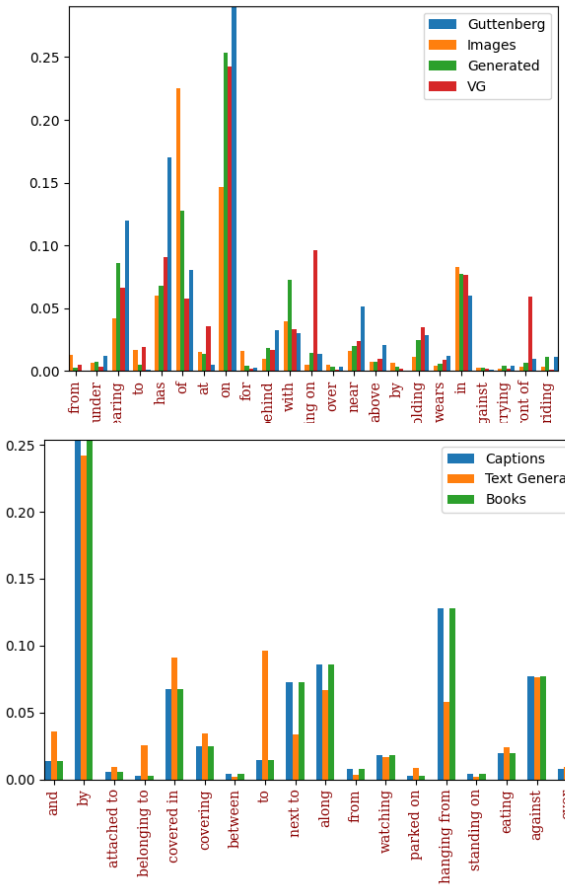

Figure 3: Top: Comparisons of the relationship distributions in the different ORM layers and VG dataset. The x-axis is the relationship label, and the y-axis is the fraction of this relationship in the ORM. This are the top-19 relationships that are in the intersection of all top-35 relationship sets. Bottom: The relationships that are in the top 35 relationships of all ORM layers but are not in the top-35 relationships of the VG dataset.

7. Implementation Details

7.1. SGG Backbone

To train the model from [7, 6] we used the code provided by [7] and [4]. TranstextNet was implemented in Pytorch on. To train Motifs and IMP We followed the protocol reported in [7]. To train VCTRee We followed the protocol reported in [4].

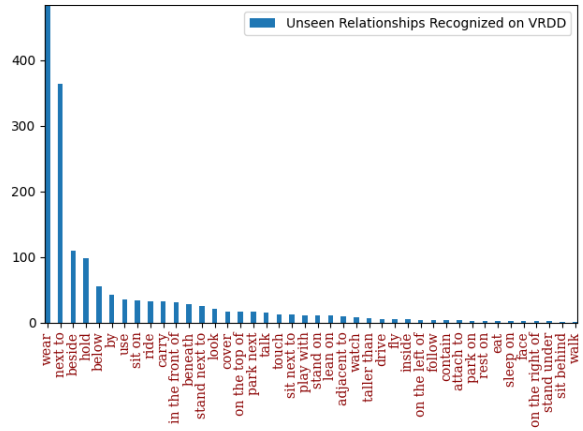

Figure 4: A histogram of ROUVR on the VRDD test set.

8. Additional Empirical Study

Due to space considerations we present additional emprical results in this document.

8.1. The Value of Different Text Sources.

We experimented with additional text sources to see how they affect performance. Table 2 presents the results for the Motifs backbone with the two additional text sources: (1) Text generated by a fine-tuned GPT-2 model [3], and (2) text from Project Gutenberg ebooks [1]. This experiment helped answer two questions: (1) How does the amount of text influence performance? (2) How does the origin of the text affect performance? On standard SGG we see that the results are similar for the various text sources. The model with generated text (GPT-2) achieves slightly better results on the mR@K metric and does better at reducing the relationship training bias. This model also outperforms the two other text models.

	Pred-CLS			
ORM	R@100	R@50	mR@100	mR@50
Captions	68.5	67	18.3	16.1
GPT-2 generated	**68.52**	**67**	**18.7**	**16.4**
Books	68.49	66.9	18.3	16.2
	Unseen Relationships			
	R@5		R@10	
Captions	16.8		23.7	
GPT-2 generated	**18.1**		**25.8**	
Books	16.2		23.1	

Table 2: Pred-Cls task. Comparison of different text sources with Motifs +TranstextNet$_A$ (both tables).

References

[1] Marie Lebert. *Le Projet Gutenberg (1971-2008)*. Project Gutenberg, 2008.

[2] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[3] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

[4] 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 Conference on Computer Vision and Pattern Recognition*, pages 6619–6628, 2019.

[5] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.

[6] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, 2017.

[7] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Conference on Computer Vision and Pattern Recognition*, 2018.