Supplemental Material - Incomparable Evaluation Summary

Rowan Zellers[1] Mark Yatskar[1,2] Sam Thomson[3] Yejin Choi[1]
[1]Paul G. Allen School of Computer Science & Engineering, University of Washington
[2]Allen Institute for Artificial Intelligence
[3]School of Computer Science, Carnegie Mellon University
{rowanz, my89, yejin}@cs.washington.edu, sthomson@cs.cmu.edu

Current work in scene graph parsing is largely inconsistent in terms of evaluation and experiments across papers are not completely comparable. In this supplementary material, we attempt to classify some of the differences and put the works together in the most comparable light.

Setup

In our paper, we compared against papers that (to the best of our knowledge) evaluated in the same way as [7]. Variation in evaluation consists of two types:

- Custom data handling, such as creating paper-specific dataset splits, changing the data pre-processing, or using different label sets.
- Omitting graph constraints, namely, allowing a head-tail pair to have multiple edge labels in system output. We hypothesize that omitting graph constraints should always lead to higher numbers, since the model is then allowed multiple guesses for challenging objects and relations.

Table 1 provides a best effort comprehensive review against all prior work that we are aware of. Other works also introduce slight variations in the tasks that are evaluated:[1]

- **Predicate Detection** (PREDDET). The model is given a list of labeled boxes, as in predicate classification, and a list of head-tail pairs that have edges in the ground truth (the model makes no edge predictions for head-tail pairs not in the ground truth).
- **Phrase Detection** (PHRDET). The model must produce a set of objects and edges, as in scene graph detection. An edge is counted as a match if the objects and predicate match the ground truth, with the IOU between the **union-boxes** of the prediction and the ground truth over 0.5 (in contrast to scene graph detection where each object box must independently overlap with the corresponding ground truth box).

Models considered

In Table 1, we list the following additional methods:

- MSDN [3]: This model is an extension of the message passing idea from [7]. In addition to using an RPN to propose boxes for objects, an additional RPN is used to propose regions for captioning. The caption generator is trained using an additional loss on the annotated regions from Visual Genome.
- MSDN-FREQ: To benchmark the performance on [3]'s split (with more aggressive preprocessing than [7] and with small objects removed), we evaluated a version of our FREQ baseline in [3]'s codebase. We took a checkpoint from the authors and replaced all edge predictions with predictions from the training set statistics from [3]'s split.
- SCR [2]: This model uses an RPN to generate triplet proposals. Messages are then passed between the head, tail, and predicate for each triplet.
- DR-NET [1]: Similar to [7], this model uses an object detector to propose regions, and then messages are passed between relationship components using an approximation to CRF inference.
- VRL [4]: This model constructs a scene graph incrementally. During training, a reinforcement learning loss is used to reward the model when it predicts correct components.
- VTE [9]: This model learns subject, predicate, and object embeddings. A margin loss is used to reward the model for predicting correct triplets over incorrect ones.
- LKD [8]: This model uses word vectors to regularize a CNN that predicts relationship triplets.

[1] We use task names from [5], despite inconsistency in whether the underlying task actually involves classification or detection.

		Graph constraints						No graph constraints									
		SGDet		SGCls		PredCls		SGDet		SGCls		PredCls		PhrDet		PredDet	
	Model	R@50	R@100	R@50	R@100	R@50	R@100	R@50	R@100	R@50	R@100	R@50	R@100	R@50	R@100	R@50	R@100
[7]'s split	VRD [5], from [7]	0.3	0.5	11.8	14.1	27.9	35.0										
	Assoc. Embed [6]	8.1	8.2	21.8	22.6	54.2	55.5	9.7	11.3	26.5	30.0	68.0	75.2				
	Message Passing [7]	3.4	4.2	21.7	24.4	44.8	53.0										
	Message Passing+	20.7	24.5	34.6	35.4	59.3	61.3	22.0	27.4	43.4	47.2	75.2	83.6	34.4	42.2	93.5	97.2
	Freq	23.5	27.6	32.4	34.0	59.9	64.1	25.3	30.9	40.5	43.7	71.3	81.2	37.2	45.0	88.3	90.1
	Freq-Overlap	26.2	30.1	32.3	32.9	60.6	62.2	28.6	34.4	39.0	43.4	75.7	82.9	41.6	49.9	94.6	96.9
	MotifNet-NoContext	26.2	29.0	34.8	35.5	63.7	65.6	29.8	34.7	43.4	46.6	78.8	85.9	43.5	50.9	94.2	97.1
	MotifNet	**27.2**	**30.3**	**35.8**	**36.5**	**65.2**	**67.1**	**30.5**	**35.8**	**44.5**	**47.7**	**81.1**	**88.3**	**44.2**	**52.1**	**96.0**	**98.4**
[3] split	MSDN [3]⋆	10.7	14.2	24.3	26.5	67.0	71.0										
	MSDN	11.7	14.0	20.9	24.0	42.3	48.2										
	MSDN-Freq	**13.5**	**15.7**	**25.8**	**27.8**	**56.0**	**61.0**										
	SCR[2]							10.67	13.81					16.58	21.54		
other split	DR-Net[1]							20.79	23.76					23.95	27.57	88.26	91.26
	VRL[4]	12.57	13.34											14.36	16.09		
	VTE[9]							5.52	6.04					9.46	10.45		
	LKD[8]															92.31	95.68

Table 1. Results with and without scene graph constraints. Horizontal lines indicate different dataset preprocessing settings (the "other split" results, to the best of our knowledge, are reported on different splits). ⋆: [3] authors acknowledge that their paper results aren't reproducible for SGCls and PredCls; their current best reproducible numbers are one line below. MSDN-Freq: Results from using node prediction from [3] and edge prediction from Freq.

Summary

The amount of variation in Table 1 requires extremely cautious interpretation. As expected, removing graph constraints significantly increases reported performance and both predicate detection and phrase detection are significantly less challenging than predicate classification and scene graph detection, respectively. On [3]'s split, the MSDN-Freq baseline outperforms MSDN on all evaluation settings, suggesting baseline is robust across alternative data settings. In total, the results suggest that our model and baselines are at least competitive with other approaches on different configurations of the task.

References

[1] B. Dai, Y. Zhang, and D. Lin. Detecting visual relationships with deep relational networks. 2017. 1, 2

[2] Y. Li, W. Ouyang, X. Wang, et al. Vip-cnn: Visual phrase guided convolutional neural network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7244–7253. IEEE, 2017. 1, 2

[3] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang. Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 1, 2

[4] X. Liang, L. Lee, and E. P. Xing. Deep Variation-structured Reinforcement Learning for Visual Relationship and Attribute Detection. *arXiv:1703.03054 [cs]*, Mar. 2017. arXiv: 1703.03054. 1, 2

[5] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, 2016. 1, 2

[6] A. Newell and J. Deng. Pixels to graphs by associative embedding. In *Advances in neural information processing systems*, 2017. 2

[7] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei. Scene Graph Generation by Iterative Message Passing. *arXiv:1701.02426 [cs]*, Jan. 2017. arXiv: 1701.02426. 1, 2

[8] R. Yu, A. Li, V. I. Morariu, and L. S. Davis. Visual relationship detection with internal and external linguistic knowledge distillation. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1, 2

[9] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua. Visual translation embedding network for visual relation detection. *CVPR*, 2017. 1, 2