Explanation of our metrics

Recall@K (R@K)

The earliest and the most widely accepted metric in scene graph generation, which is firstly adopted by Visual relationship detection with language priors. Since the ground-truth annotations of relationships are incomplete, it's improper to use simple accurary as the metric. Therefore, Lu et al. transfer it to a retrieve-like problem: the relationships are not only required to be correctly classified, but also required to have as higher score as possible, so they can be retrieved from plenty of 'none' relationship pairs.

No Graph Constraint Recall@K (ng-R@K)

It's firstly used by Pixel2Graph and named by Neural-MOTIFS. The former paper significantly improves the R@K results by allowing each pair to have multiple predicates, which means for each subject-object pair, all the 50 predicates will be involved in the recall ranking not just the one with highest score. Since predicates are not exclusive, 'on' and 'riding' can both be correct. This setting significantly improves the R@K. To fairly compare with other methods, Neural-MOTIFS named it as the No Graph Constraint Recall@K (ngR@K).

Mean Recall@K (mR@K)

It is proposed by our work VCTree and Chen et al.s'KERN at the same time (CVPR 2019), although we didn't make it as our main contribution and only listed the full results on the supplementary material. However, we also acknowledge the contribution of KERN, for they gave more mR@K results of previous methods. The main motivation of Mean Recall@K (mR@K) is that the VisualGenome dataset is biased towards dominant predicates. If the 10 most frequent predicates are correctly classified, the accuracy would reach 90% even the rest 40 kinds of predicates are all wrong. This is definitely not what we want. Therefore, Mean Recall@K (mR@K) calculates Recall@K for each predicate category independently then report their mean.

No Graph Constraint Mean Recall@K (ng-mR@K)

The same mean Recall metric, but for each pair of objects, all possible predicates are valid candidates (the original mean Recall@K only considers the predicate with maximum score of each pair as the valid candidate to calculate Recall).

Zero Shot Recall@K (zR@K)

It is firstly used by Visual relationship detection with language priors for VRD dataset, and firstly reported by Unbiased Scene Graph Generation from Biased Training for VisualGenome dataset. In short, it only calculates the Recall@K for those subject-predicate-object combinations that not occurred in the training set.

No Graph Constraint Zero Shot Recall@K (ng-zR@K)

The same zero-shot Recall metric, but for each pair of objects, all possible predicates are valid candidates (the original zero-shot Recall@K only considers the predicate with maximum score of each pair as the valid candidate to calculate Recall).

Top@K Accuracy (A@K)

It is actually caused by the misunderstanding of PredCls and SGCls protocols. Contrastive Losses reported Recall@K of PredCls and SGCls by not just giving ground-truth bounding boxes, but also giving the ground-truth subject-object pairs, so no ranking is involved. The results can only be considerred as Top@K Accuracy (A@K) for the given K ground-truth subject-object pairs.

Sentence-to-Graph Retrieval (S2G)

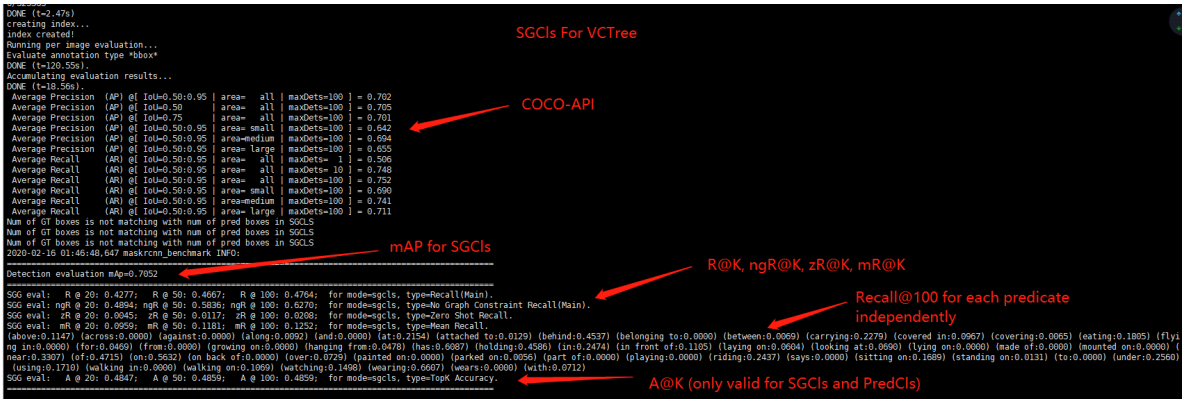
S2G is proposed by Unbiased Scene Graph Generation from Biased Training as an ideal downstream task that only relies on the quality of SGs, for the existing VQA and Captioning are too complicated and challenged by their own bias. It takes human descriptions as queries, searching for matching scene graphs (images), where SGs are considered as the symbolic representations of images. More details will be explained in S2G-RETRIEVAL.md.

Two Common Misunderstandings in SGG Metrics

When you read/follow a SGG paper, and you find that its performance is abnormally high for no obvious reasons, whose authors could mess up some metrics.

1. Not differentiate Graph Constraint Recall@K and No Graph Constraint Recall@K. The setting of With/Without Constraint is introduced by Neural-MOTIFS. However, some early work and a few recent researchers don't differentiate these two setting, using No Graph Constraint results to compare with previous work With Graph Constraint. TYPICAL SYMPTOMS: 1) Recall@100 of PredCls is larger than 75%, 2) not mention With/Without Graph Constraint in the original paper. TYPICAL Paper:Pixel2Graph (Since this paper is published before MOTIFS, they didn't mean to take this advantage, and they are actually the fathers of No Graph Constraint setting while MOTIFS is the one who named this baby.)
2. Some researchers misunderstand the protocols of PredCls and SGCls. These two protocols only give ground-truth bounding boxes NOT ground-truth subject object pairs. Some works only predict relationships for ground-truth subject-object pairs in PredCls and SGCls, so their PredCls and SGCls results will extremely high. Note that Recall@K metric is a ranking metric, using ground-truth subject-object pairs can be considerred as giving the perfect ranking. In order to separate from normal PredCls and SGCls, I name this kind of setting as Top@K Accuracy, which is only applicable to PredCls and SGCls. TYPICAL SYMPTOMS: 1) results of PredCls and SGCls are extremely high while results of SGGen are normal, 2) Recall@50 and Recall@100 of PredCls and SGCls are exactly the same, since the ranking is perfect (Recall@20 is less, for some images have groud-truth relationships more than 20). TYPICAL Paper:Contrastive Losses.

Output Format of Our Code



Reported Results

The results of reimplemented IMP, MOTIFS, VCTree and our Transformer with X-101-FPN backbone

All the following results only used two 1080ti GPUs with batch size 8 (increasing batch size will further improves some models a little bit). Note that the reimplemented VCTree is not exactly the same as the original work. It's an optimized version for SGCls and SGGen. But PredCls seems not as good as previous, I will try to find the reason later. Hybrid Learning is discarded for simplicity.

Recall@K

Models	SGGen R@20	SGGen R@50	SGGen R@100	SGCls R@20	SGCls R@50	SGCls R@100	PredCls R@20	PredCls R@50	PredCls R@100
IMP	18.09	25.94	31.15	34.01	37.48	38.50	54.34	61.05	63.06
MOTIFS	25.48	32.78	37.16	35.63	38.92	39.77	58.46	65.18	67.01
Transformer	25.55	33.04	37.40	36.87	40.18	41.02	59.06	65.55	67.29
VCTree	24.53	31.93	36.21	42.77	46.67	47.64	59.02	65.42	67.18

No Graph Constraint Recall@K

Models	SGGen ngR@20	SGGen ngR@50	SGGen ngR@100	SGCls ngR@20	SGCls ngR@50	SGCls ngR@100	PredCls ngR@20	PredCls ngR@50	PredCls ngR@100
IMP	18.35	27.02	33.89	38.70	46.78	51.20	62.14	76.82	84.97
MOTIFS	27.04	36.58	43.43	40.58	48.48	51.98	66.39	81.02	88.24
Transformer	27.14	36.98	43.90	42.31	50.18	53.93	67.45	81.83	88.95
VCTree	26.14	35.73	42.34	48.94	58.36	62.70	67.20	81.63	88.83

Zero Shot Recall@K

Note: IMP achieves highest Zero Shot Recall@K because it doesn't include any explicit or implicit object label embeddings for predicate prediction.

Models	SGGen zR@20	SGGen zR@50	SGGen zR@100	SGCls zR@20	SGCls zR@50	SGCls zR@100	PredCls zR@20	PredCls zR@50	PredCls zR@100
IMP	0.18	0.38	0.77	2.01	3.30	3.92	12.17	17.66	20.25
MOTIFS	0.0	0.05	0.11	0.32	0.68	1.13	1.08	3.24	5.36
Transformer	0.04	0.14	0.29	0.34	0.91	1.39	1.35	3.63	5.64
VCTree	0.10	0.31	0.69	0.45	1.17	2.08	1.04	3.27	5.51

Mean Recall@K

Models	SGGen mR@20	SGGen mR@50	SGGen mR@100	SGCls mR@20	SGCls mR@50	SGCls mR@100	PredCls mR@20	PredCls mR@50	PredCls mR@100
IMP	2.75	4.17	5.30	5.21	6.18	6.53	8.85	10.97	11.77
MOTIFS	4.98	6.75	7.90	6.68	8.28	8.81	11.67	14.79	16.08
Transformer	6.01	8.13	9.56	8.14	10.09	10.73	12.77	16.30	17.63
VCTree	5.38	7.44	8.66	9.59	11.81	12.52	13.12	16.74	18.16

Top@K Accuracy

Models	SGGen A@20	SGGen A@50	SGGen A@100	SGCls A@20	SGCls A@50	SGCls A@100	PredCls A@20	PredCls A@50	PredCls A@100
IMP	-	-	-	39.19	39.30	39.30	64.88	65.12	65.12
MOTIFS	-	-	-	40.41	40.50	40.50	68.87	69.14	69.14
Transformer	-	-	-	41.75	41.84	41.84	69.08	69.36	69.36
VCTree	-	-	-	48.47	48.59	48.59	68.92	69.19	69.19

The results of Unbiased Scene Graph Generation from Biased Training with X-101-FPN backbone

Note that if you are using the default VCTree settings of this project, all results of VCTree should be better than what we reported in Unbiased Scene Graph Generation from Biased Training, i.e., the following results, because we optimized the tree construction network after the publication.

Recall@K and Mean Recall@K

			Predicate Classification		Scene Graph Classification		Scene Graph Detection	
Model	Fusion	Method	R@20 / 50 / 100	mR@20 / 50 / 100	R@20 / 50 / 100	mR@20 / 50 / 100	R@20 / 50 / 100	mR@20 / 50 / 100
IMP+ [60, 6]	-	-	52.7 / 59.3 / 61.3	- / 9.8 / 10.5	31.7 / 34.6 / 35.4	- / 5.8 / 6.0	14.6 / 20.7 / 24.5	- / 3.8 / 4.8
FREQ [65, 51]	-	-	53.6 / 60.6 / 62.2	8.3 / 13.0 / 16.0	29.3 / 32.3 / 32.9	5.1 / 7.2 / 8.5	20.1 / 26.2 / 30.1	4.5 / 6.1 / 7.1
MOTIFS [65, 51]	-	-	58.5 / 65.2 / 67.1	10.8 / 14.0 / 15.3	32.9 / 35.8 / 36.5	6.3 / 7.7 / 8.2	21.4 / 27.2 / 30.3	4.2 / 5.7 / 6.6
KERN [6]	-	-	- / 65.8 / 67.6	- / 17.7 / 19.2	- / 36.7 / 37.4	- / 9.4 / 10.0	- / 27.1 / 29.8	- / 6.4 / 7.3
VCTree [51]	-	-	60.1 / 66.4 / 68.1	14.0 / 17.9 / 19.4	35.2 / 38.1 / 38.8	8.2 / 10.1 / 10.8	22.0 / 27.9 / 31.3	5.2 / 6.9 / 8.0
MOTIFS†	SUM	Baseline	59.5 / 66.0 / 67.9	11.5 / 14.6 / 15.8	35.8 / 39.1 / 39.9	6.5 / 8.0 / 8.5	25.1 / 32.1 / 36.9	4.1 / 5.5 / 6.8
		Focal	59.2 / 65.8 / 67.7	10.9 / 13.9 / 15.0	36.0 / 39.3 / 40.1	6.3 / 7.7 / 8.3	24.7 / 31.7 / 36.7	3.9 / 5.3 / 6.6
		Reweight	45.4 / 57.0 / 61.7	16.0 / 20.0 / 21.9	24.2 / 29.5 / 31.5	8.4 / 10.1 / 10.9	18.3 / 24.4 / 29.3	6.5 / 8.4 / 9.8
		Resample	57.6 / 64.6 / 66.7	14.7 / 18.5 / 20.0	34.5 / 37.9 / 38.8	9.1 / 11.0 / 11.8	23.2 / 30.5 / 35.4	5.9 / 8.2 / 9.7
		X2Y	58.3 / 65.0 / 66.9	13.0 / 16.4 / 17.6	35.2 / 38.6 / 39.5	6.9 / 8.6 / 9.2	24.8 / 32.1 / 36.7	5.1 / 6.9 / 8.1
		X2Y-Tr	59.0 / 65.3 / 66.9	11.6 / 14.9 / 16.0	35.5 / 38.9 / 39.7	6.5 / 8.4 / 9.1	25.5 / 32.8 / 37.2	5.0 / 6.9 / 8.1
		TE	34.3 / 46.7 / 51.7	18.2 / 25.3 / 29.0	25.5 / 32.5 / 35.4	8.1 / 12.0 / 14.0	14.8 / 20.1 / 23.9	5.7 / 8.0 / 9.6
		NIE	0.6 / 1.0 / 1.3	0.6 / 1.1 / 1.4	28.6 / 35.0 / 37.4	6.1 / 9.0 / 10.6	17.3 / 22.7 / 26.8	3.8 / 5.1 / 6.0
		TDE	33.6 / 46.2 / 51.4	18.5 / 25.5 / 29.1	21.7 / 27.7 / 29.9	9.8 / 13.1 / 14.9	12.4 / 16.9 / 20.3	5.8 / 8.2 / 9.8
	GATE	Baseline	58.9 / 65.5 / 67.4	12.2 / 15.5 / 16.8	36.2 / 39.4 / 40.1	7.2 / 9.0 / 9.5	25.8 / 33.3 / 37.8	5.2 / 7.2 / 8.5
		TDE	38.7 / 50.8 / 55.8	18.5 / 24.9 / 28.3	21.8 / 27.2 / 29.5	11.1 / 13.9 / 15.2	5.9 / 7.4 / 8.4	6.6 / 8.5 / 9.9
VTransE†	SUM	Baseline	59.0 / 65.7 / 67.6	11.6 / 14.7 / 15.8	35.4 / 38.6 / 39.4	6.7 / 8.2 / 8.7	23.0 / 29.7 / 34.3	3.7 / 5.0 / 6.0
		TDE	36.9 / 48.5 / 53.1	17.3 / 24.6 / 28.0	19.7 / 25.7 / 28.5	9.3 / 12.9 / 14.8	13.5 / 18.7 / 22.6	6.3 / 8.6 / 10.5
	GATE	Baseline	58.7 / 65.3 / 67.1	13.6 / 17.1 / 18.6	34.6 / 38.1 / 38.9	6.6 / 8.2 / 8.7	24.5 / 31.3 / 35.5	5.1 / 6.8 / 8.0
		TDE	40.0 / 50.7 / 54.9	18.9 / 25.3 / 28.4	23.0 / 28.8 / 31.1	9.8 / 13.1 / 14.7	13.7 / 19.0 / 22.9	6.0 / 8.5 / 10.2
VCTree†	SUM	Baseline	59.8 / 66.2 / 68.1	11.7 / 14.9 / 16.1	37.0 / 40.5 / 41.4	6.2 / 7.5 / 7.9	24.7 / 31.5 / 36.2	4.2 / 5.7 / 6.9
		TDE	36.2 / 47.2 / 51.6	18.4 / 25.4 / 28.7	19.9 / 25.4 / 27.9	8.9 / 12.2 / 14.0	14.0 / 19.4 / 23.2	6.9 / 9.3 / 11.1
	GATE	Baseline	59.1 / 65.5 / 67.4	12.4 / 15.4 / 16.6	35.4 / 38.9 / 39.8	6.3 / 7.5 / 8.0	24.8 / 31.8 / 36.1	4.9 / 6.6 / 7.7
		TDE	39.1 / 49.9 / 54.5	17.2 / 23.3 / 26.6	22.8 / 28.8 / 31.2	8.9 / 11.8 / 13.4	14.3 / 19.6 / 23.3	6.3 / 8.6 / 10.3

Zero Shot Recall@K

Zero-Shot Relationship Retrieval			PredCls	SGCls	SGDet
Model	Fusion	Method	R@50/100	R@50/100	R@50/100
MOTIFS†	SUM	Baseline	10.9 / 14.5	2.2 / 3.0	0.1 / 0.2
		Focal	10.9 / 14.4	2.2 / 3.1	0.1 / 0.3
		Reweight	0.7 / 0.9	0.1 /0.1	0.0 / 0.0
		Resample	11.1 / 14.3	2.3 / 3.1	0.1 / 0.3
		X2Y	11.8 / 17.6	2.3 / 3.7	1.6 / 2.7
		X2Y-Tr	13.7 / 17.6	3.1 / 4.2	1.8 / 2.8
		TE	14.2 / 18.1	1.4 / 2.0	1.4 / 1.8
		NIE	2.4 / 3.2	0.2 / 0.4	0.3 / 0.6
		TDE	**14.4 / 18.2**	**3.4 / 4.5**	**2.3 / 2.9**
	GATE	Baseline	7.4 / 10.6	0.9 / 1.3	0.2 / 0.4
		TDE	**7.7 / 11.0**	**1.9 / 2.6**	**1.9 / 2.5**
VTransE†	SUM	Baseline	11.3 / 14.7	2.5 / 3.3	0.8 / 1.5
		TDE	**13.3 / 17.6**	**2.9 / 3.8**	**2.0 / 2.7**
	GATE	Baseline	4.2 / 5.9	1.9 / 2.6	**1.9** / 2.6
		TDE	**5.3 / 7.9**	**2.1 / 3.0**	**1.9 / 2.7**
VCTree†	SUM	Baseline	10.8 / 14.3	1.9 / 2.6	0.2 / 0.7
		TDE	**14.3 / 17.6**	**3.2 / 4.0**	**2.6 / 3.2**
	GATE	Baseline	4.4 / 6.8	2.5 / 3.3	1.8 / 2.7
		TDE	**5.9 / 8.1**	**3.0 / 3.7**	**2.2 / 2.8**

Table 2. The results of Zero-Shot Relationship Retrieval.