Probabilistic Debiasing of Scene Graphs

Bashirul Azam Biswas (*biswab@rpi.edu*)
Qiang Ji (*jiq@rpi.edu*)
Dept. of ECSE, RPI

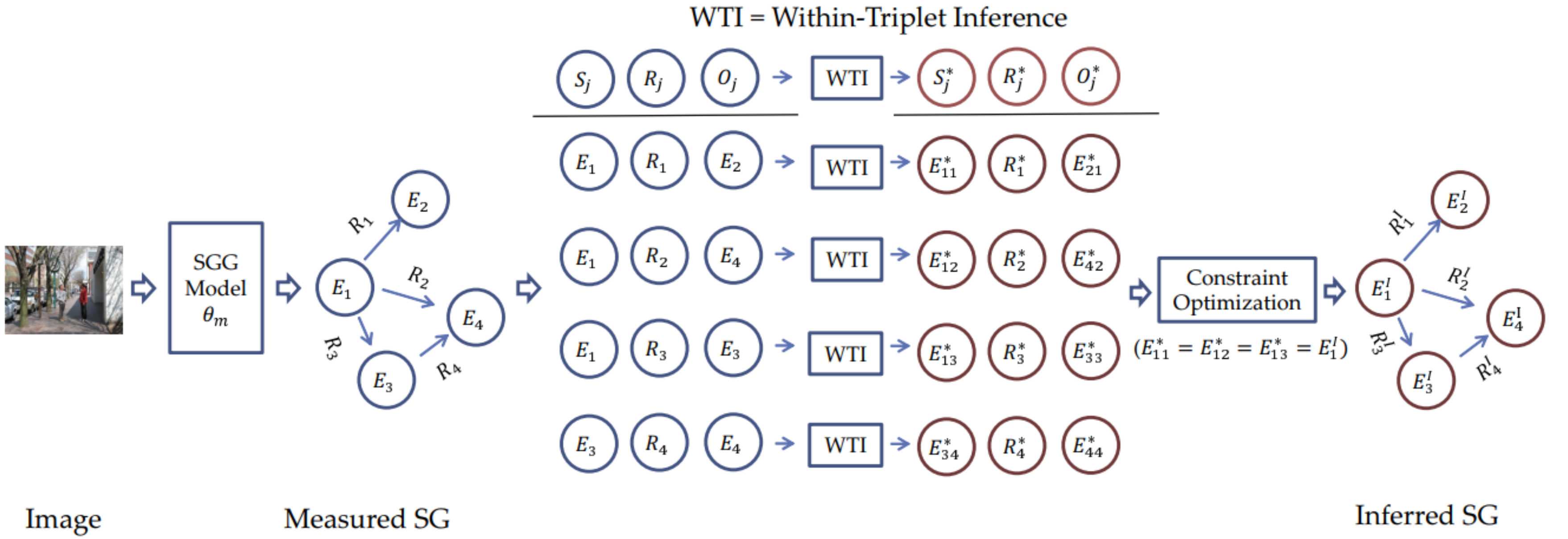
Paper ID : 9960

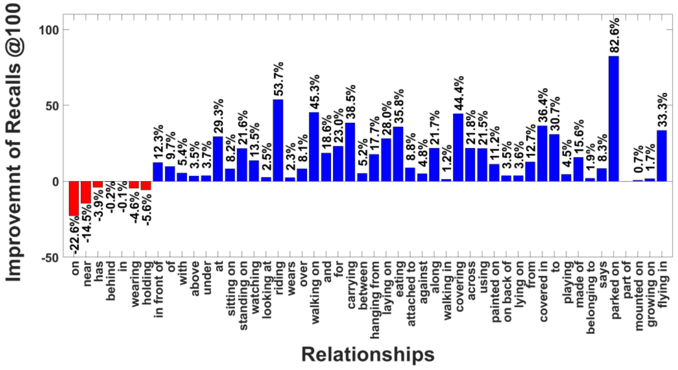
Poster Tag : WED-AM-210

Preview of Our Work

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

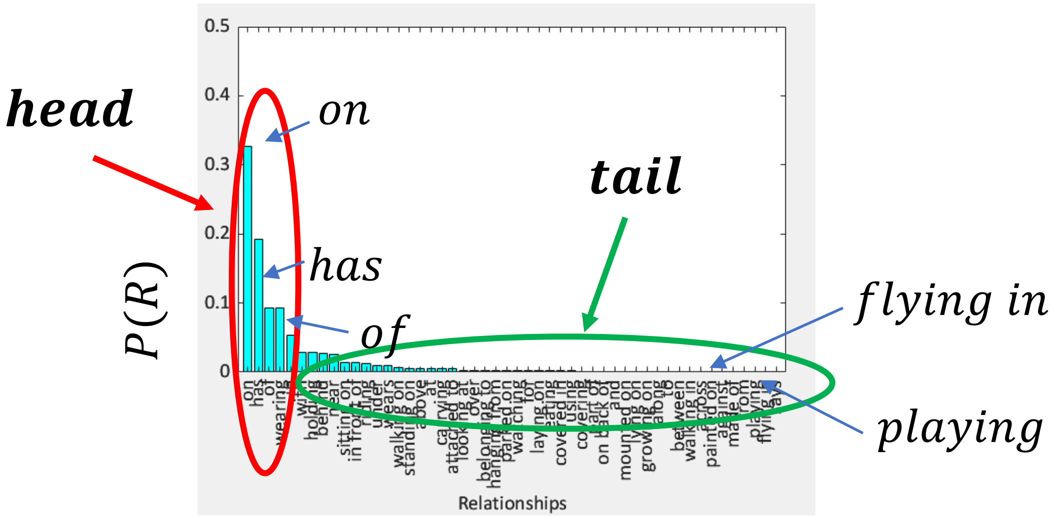
- We ***debias*** predicted scene graph triplets with ***within-triplet*** Bayesian network.



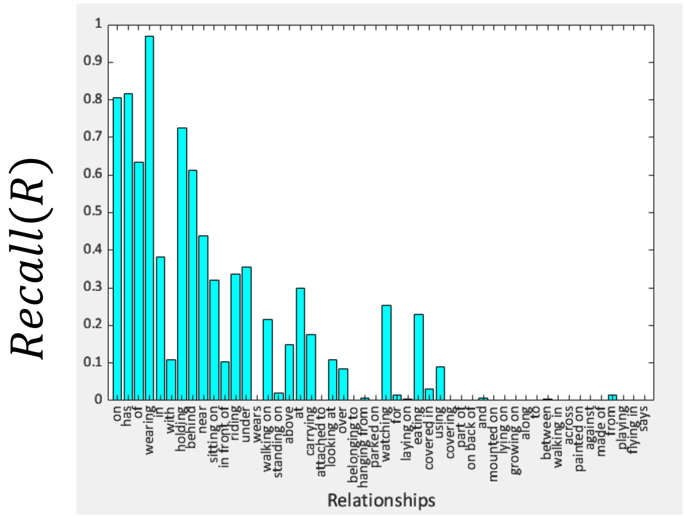


- We improve the performance of ***tail*** classes at the minimal expense of ***head*** classes

Why debias Scene Graphs?



Long-tailed distribution of relationship labels



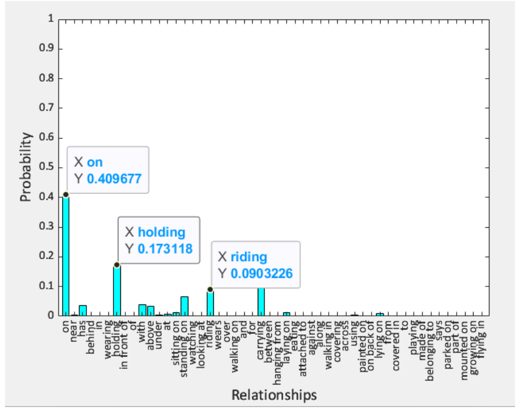
Poor performance of tail classes in SOTA model

- ❑ Deep learning-based Scene Graph Generation (SGG) models perform poorly on the tail classes
- ❑ Traditional debiasing schemes
  - ❑ improve the ***tail*** classes with significant hurting of the ***head*** classes
  - ❑ ignore within-triplet prior
- ❑ We debias scene graph with restoring within-triplet prior and hurt the ***head*** classes minimally

Within-triplet Prior in Scene Graphs

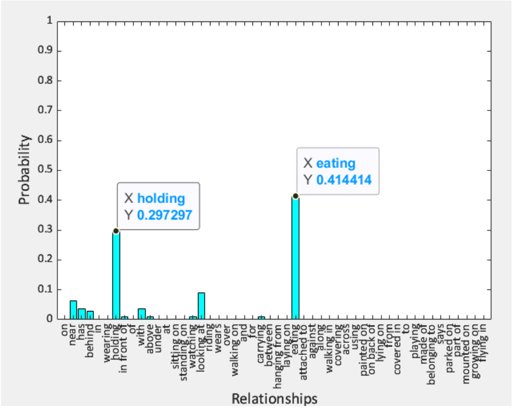
- A ‘**man**’ will most likely be ‘**on**’ or ‘**hold**’ a ‘**surfboard’**.

$P(R|\ S = man, O = surfboard)$



- A ‘**man**’ will most likely ‘**eat**’ or ‘**hold**’ a ‘**pizza’**.

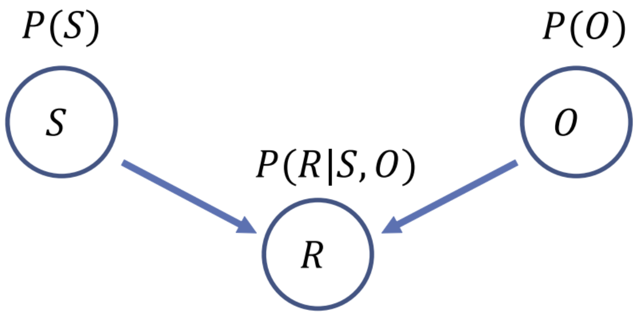
$P(R|\ S = man, O = pizza)$



- Distribution of relationship is strongly dependent on its subject and object !

Bayesian Network (BN) to capture within-triplet prior

- Joint distribution of subject, relationship, and object is denoted by $\boldsymbol{P(S, R, O)}$
- We aim to capture this joint distribution with a Bayesian Network
- Assumptions -
  - Relationship is dependent both on its subject and object,
  - Subject and object are **independent** of each other,
  - Subject and object become **dependent** given the relationships.



- Under these assumptions –
  - $P(S, R, O) = P(R|S, O)P(S)P(O)$

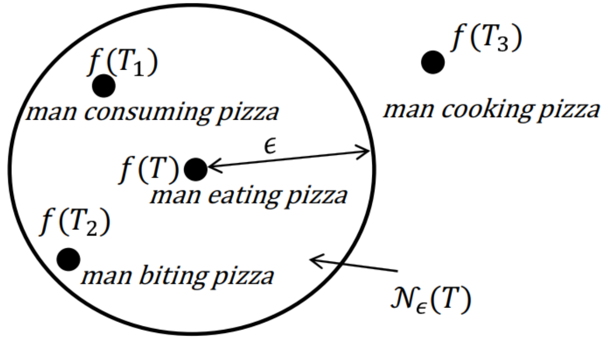
Learning Within-Triplet Bayesian Network

- Learning with annotated triplets

$$P(R=r|S=s,O=o) = \frac{N^c_{s,r,o}}{\sum_{r'} N^c_{s,r',o}} \longrightarrow \text{Learning } P(R|S,O)$$

$$P(R) = \sum_{S,O} P(R,S,O) \longrightarrow \text{Learning } P(R)$$

- Learning with augmented triplets
  - Top-50 relationships from full dataset are chosen for SGG task.
  - Many other relationships in the dataset outside these top-50 bear similar meaning
    - man-consuming-pizza is similar to man-eating-pizza
  - We augment triplet counts with similar triplets
  - Similarity is calculated in embedding space of triplets.

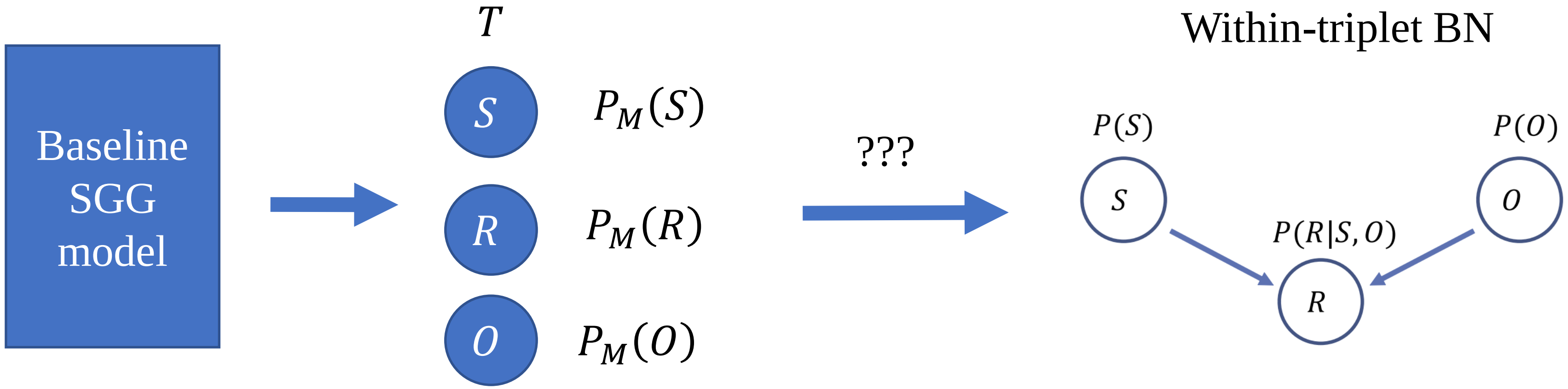


$$N^a_{s,r,o} = \begin{cases} N^c_{s,r,o} + \sum_{T_i \in \mathcal{N}_\epsilon(T)} N_{s,r_i,o} \\ N^c_{s,r,o} \ \text{ if } \mathcal{N}_\epsilon(T) = \emptyset \end{cases} \longrightarrow \text{Augmented count}$$

$$\mathcal{N}_\epsilon(T) = \{T_i : \phi(f(T), f(T_i)) < \epsilon\}$$

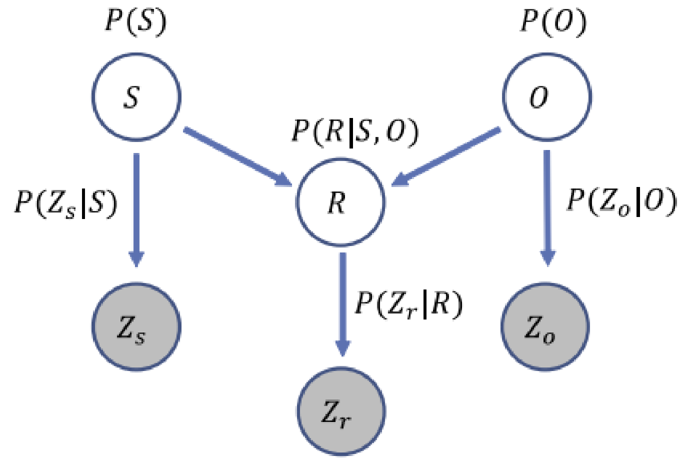
Uncertain Evidence of Triplets

- A baseline measurement model produces probability of –
  - subject (S), relationship (R), and object (O) of each triplet T in a scene graph.
- We denote these probabilities as $P_M(S), P_M(O),$ and $P_M(R)$.
- We incorporate these probabilities into our proposed BN as uncertain evidence to perform posterior inference.
- Uncertain evidence is incorporated as virtual evidence.



Virtual Evidence of Within-Triplet BN

- Three virtual evidence nodes Z_s, Z_o, Z_r are created as child of their respective parents S, O, and R.
- The conditional probabilities of these nodes are specified from their likelihood ratios
  - Likelihood ratio is obtained by scaling the biased measured probability by the biased marginal probability
  - The scaling bolsters the probability of ***tail*** classes of ***head***-driven baseline model.



Virtual Evidence (VE) of triplet elements

Measured probability

$$P(Z_s=1|s_1):..:P(Z_s=1|s_n)=\frac{P_M(s_1)}{P(s_1)}:..:\frac{P_M(s_n)}{P(s_n)}$$

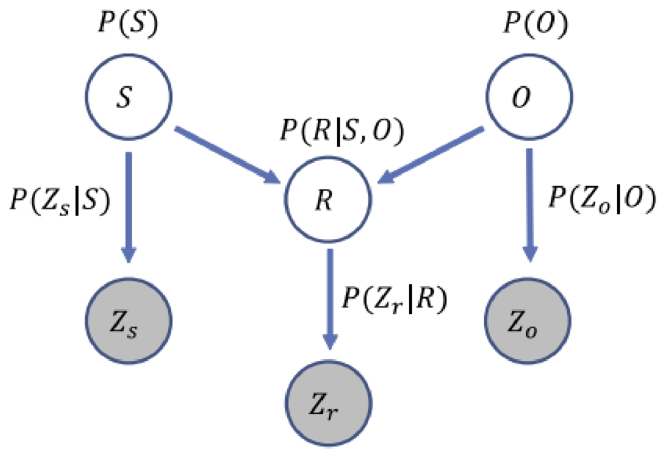
Marginal probability

$$P(Z_o=1|o_1):..:P(Z_o=1|o_n)=\frac{P_M(o_1)}{P(o_1)}:..:\frac{P_M(o_n)}{P(o_n)}$$

$$P(Z_r=1|r_1):..:P(Z_r=1|r_n)=\frac{P_M(r_1)}{P(r_1)}:..:\frac{P_M(r_n)}{P(r_n)}$$

Specifying conditional probability of VE nodes

Posterior inference of Within-Triplet BN



$$P(S,R,O|Z_s=1,Z_o=1,Z_r=1)$$

$$\propto P(Z_s=1|S)P(S)P(Z_o=1|O)P(O)P(Z_r=1|R)P(R|S,O)$$

$$=P_M(S)P_M(O)\frac{P_M(R)}{P(R)}P(R|S,O)$$

$$S^*,R^*,O^*=\arg\max_{S,R,O}P(S,R,O|Z_s=1,Z_o=1,Z_r=1)$$

$$=\arg\max_{S,R,O}P_M(S)P_M(O)\frac{P_M(R)}{P(R)}P(R|S,O)$$

Measured subject prob.

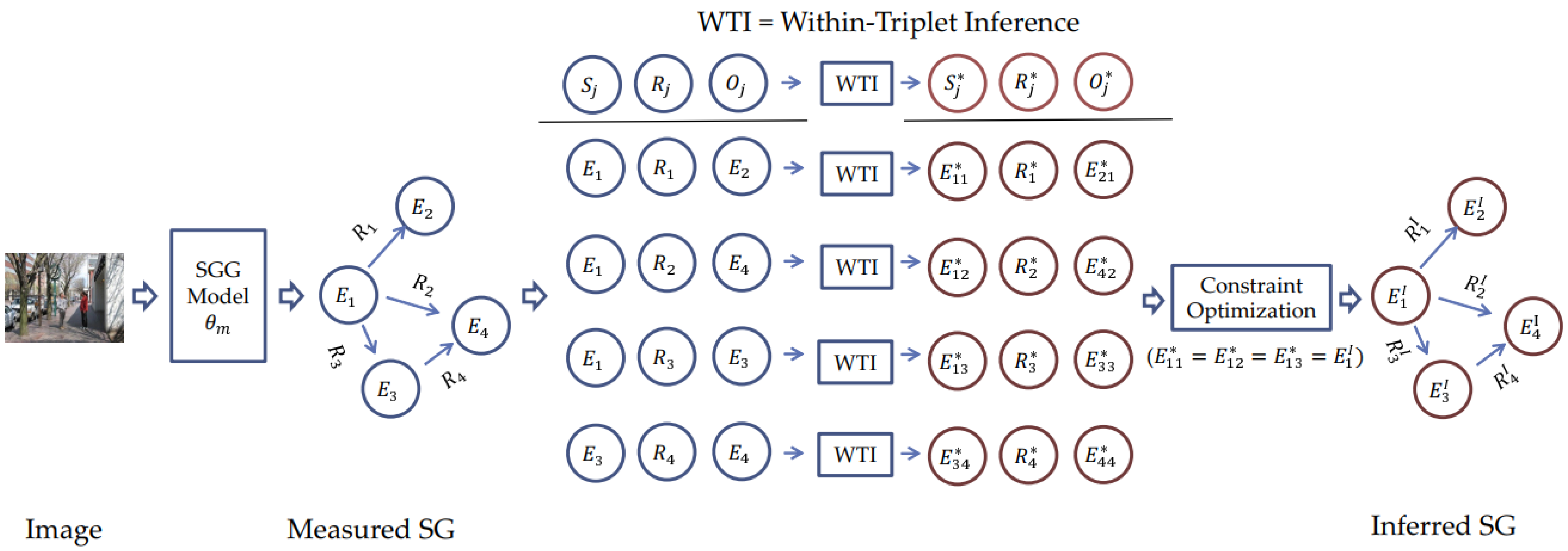
Measured object prob.

Measured relationship prob.

Conditional relationship prob.

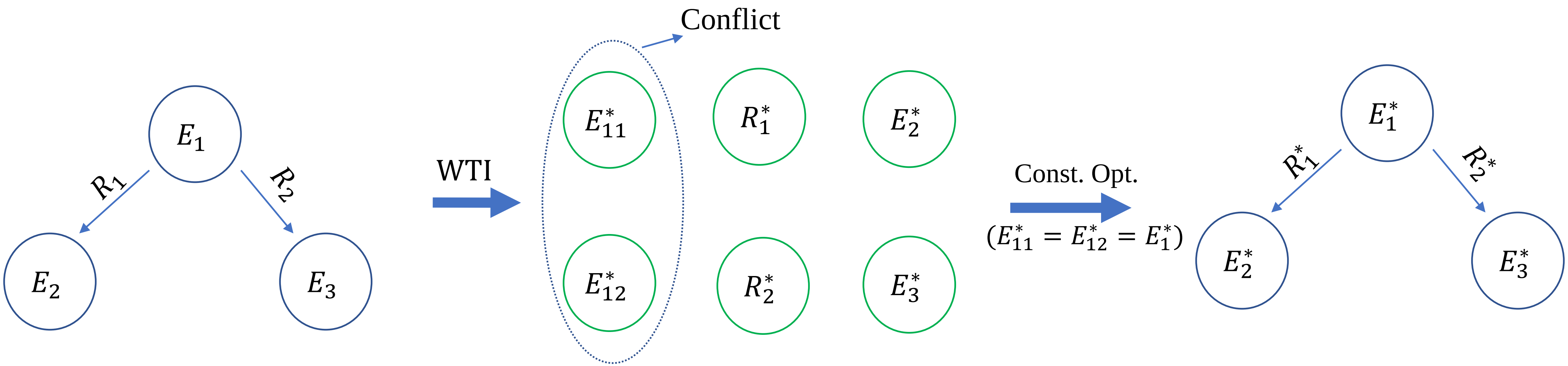
Marginal relationship prob.

Overview of our proposed approach



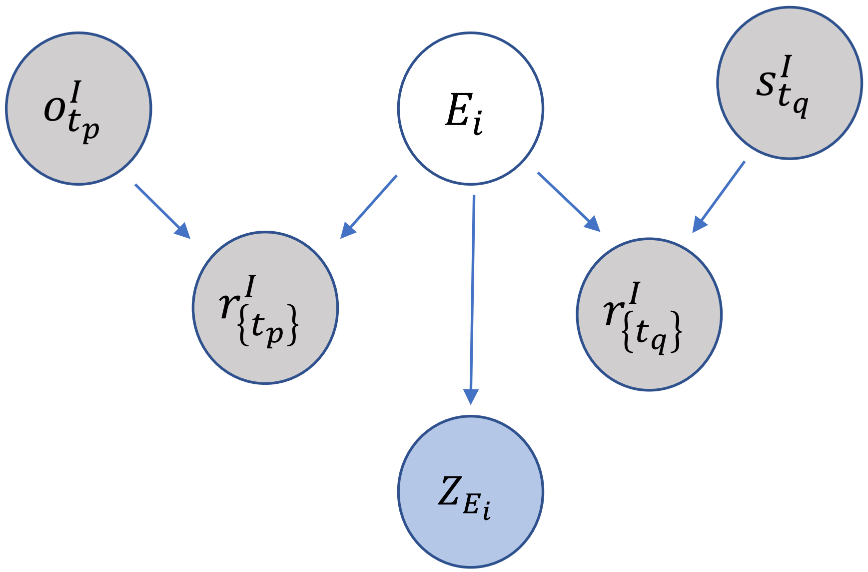
Constrained Optimization for Adjusting Inference Results

- Subject or object of any triplet may be shared by other triplets.
- Therefore, posterior inference of individual triplet may produce different results for the same subject or object.
- Need to perform a constrained optimization to resolve such conflicts.

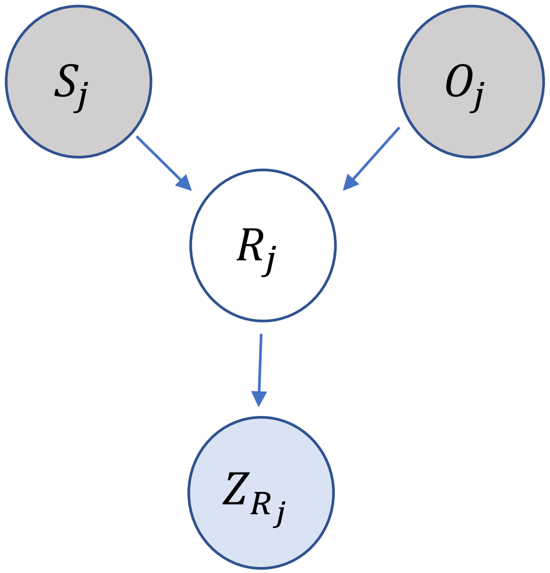


Constrained Optimization for Adjusting Inference Results

- Two-step iterative optimization. At each iteration, we perform -
  - **Object updating**: refine each object node separately keeping all other nodes fixed.
  - **Relationship updating**: refine each relationship node separately keeping all other nodes fixed.



Object Updating



Relationship Updating

Evaluation Metric

- **Recall $R@K$**

  - M_I → Total matched triplets in image I in top-K predicted triplets
  - G_I → Total ground truth triplets in image I

  - $R@K = \frac{1}{N_I} \sum_I \frac{M_I}{G_I}$

- **Mean Recall $mR@K$**

  - $M_{I,R}$ → Total matched triplets of relation R in image I in top-K predicted triplets
  - $G_{I,R}$ → Total ground truth triplets of relation R in image I

  - $mR@K = \frac{1}{N_R} \sum_R \frac{1}{N_I} \sum_I \frac{M_{I,R}}{G_{I,R}}$

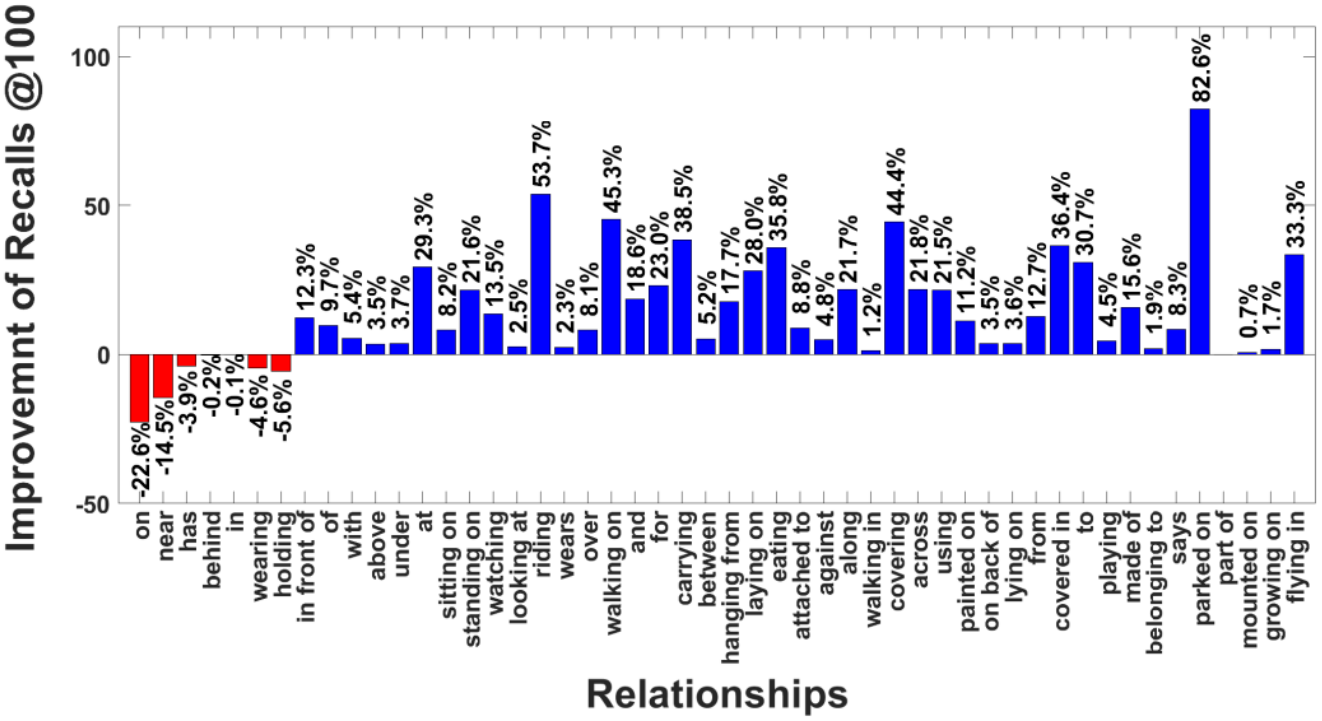
Performance Comparison from Baseline Model

- Recall is ***decreasing***
- Mean Recall is ***increasing***

DS	Method	Recall and Mean Recall @K					
		PredCls		SGCls		SGDet	
		R@50/100	mR@50/100	R@50/100	mR@50/100	R@50/100	mR@50/100
VG	VCTree$^\diamond$ [29]	65.46/ 67.18	15.36/ 16.61	44.15/ 45.11	9.17/ 9.83	29.94/ 32.57	6.21/ 6.96
	Inf-VCTree	59.50/ 60.97 ↓	**28.14/ 30.72** ↑	40.69/ 41.55 ↓	**17.31/ 19.40** ↑	27.74/ 30.10 ↓	**10.40/ 11.86** ↑
GQA	VCTree$^\diamond$ [29]	68.83/ 70.14	22.07/ 23.01	35.04/ 35.58	10.59/ 10.97	27.21/ 28.79	7.03/ 7.75
	Inf-VCTree	62.80/ 64.05 ↓	**39.44/ 41.63** ↑	32.23/ 32.80 ↓	**19.18/ 20.03** ↑	25.10/ 26.45 ↓	**13.57/ 15.12** ↑

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Performance Comparison from Baseline Model



- Relationships are ordered with descending order of their frequencies
- *head* classes such as ‘on’, ‘near’ , ‘has’ , ‘behind’ is dropping
- *tail* classes are improving
- Typical behavior in SGG debiasing work

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Performance Comparison with SOTA debiasing methods

(Baseline) → VCTree

Method	Re-train	R@K @50/100	mR@K @50/100
VCTree [29]	-	65.5/ 67.2	15.4/ 16.6
Unb-VCTree [28]	**No**	47.2/ 51.6	25.4/ 28.7
DLFE-VCTree [4]	**Yes**	51.8/ 53.5	25.3/ 27.1
NICE-VCTree [13]	**Yes**	55.0/ 56.9	**30.7/ 33.0**
Inf-VCTree (Ours)	**No**	**59.5/ 61.0**	28.1/ 30.7

- Other debiasing methods
  - decrease the $R@K$ significantly since they do not incorporate the within-triplet prior.
  - require re-training of the baseline models.
- Our method
  - hurts the $R@K$ less brutally.
  - requires no re-training of the baseline model.

Conclusion

- We debiased the predicted scene graphs with minimal hurting of the ***head*** classes
- We incorporated within-triplet prior in debiasing step through a Bayesian Network
- Triplet evidence is incorporated into BN with virtual evidence
- Possible conflicts in subject and object are resolved with a constraint optimization step
- Our method
  - improves the $\boldsymbol{tail}$ classes with minimal hurting of the ***head*** classes
  - requires no re-training of the baseline models
  - can be incorporated as a plug-and-play module