Single-Stage Visual Relationship Learning using Conditional Queries

Alakh Desai[1], Tz-Ying Wu[1], Subarna Tripathi[2], Nuno Vasconcelos[1]

[1]University of California San Diego, USA [2]Intel Labs, USA

Introduction

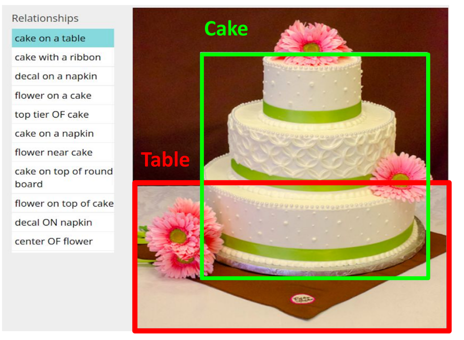
Scene graphs provide a structured description of a scene

Prior methods perform either:

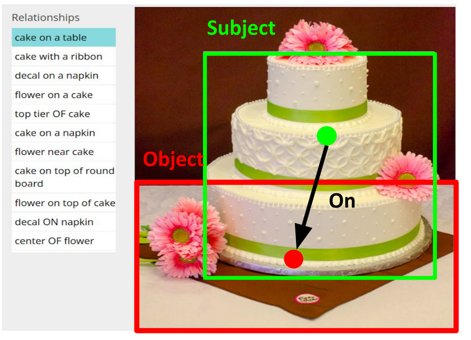
1. **Entity first prediction**
   a. is combinatorially expensive
   b. does not capture interaction features well
2. **Single shot set prediction**
   a. is a multi-task learning task
   b. achieves poor performance overall

Our predicate first approach

- decouples the multi-task problem
- does not require entity pair matching
- learns interaction features better

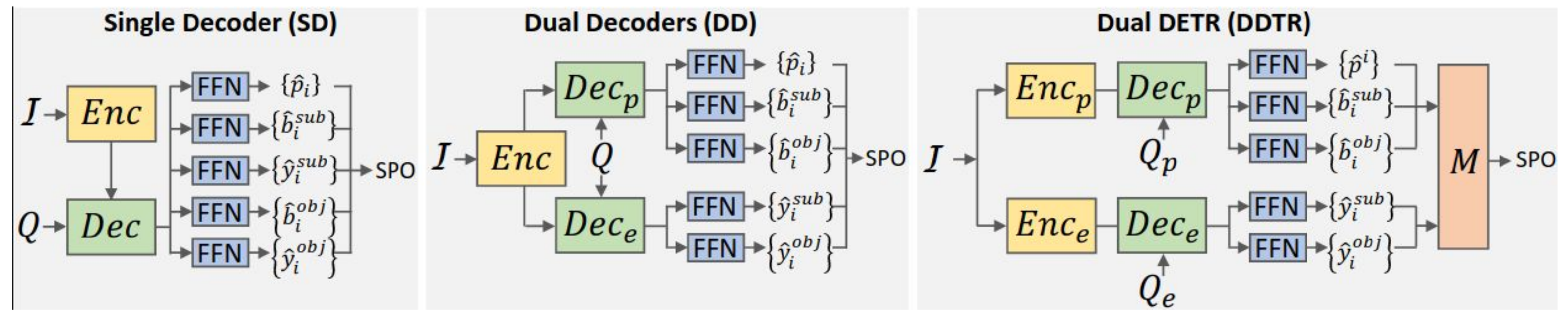


Entity first SGG



Predicate first SGG

Baseline Architectures



Baseline architectures

1. **Single Decoder (SD):** uses a pair of encoder and decoder modules and 5 FFNs to decode each element < (b_{sub}, y_{sub}) − p − (b_{obj} , y_{obj}) >. This architecture promotes maximum entanglement.

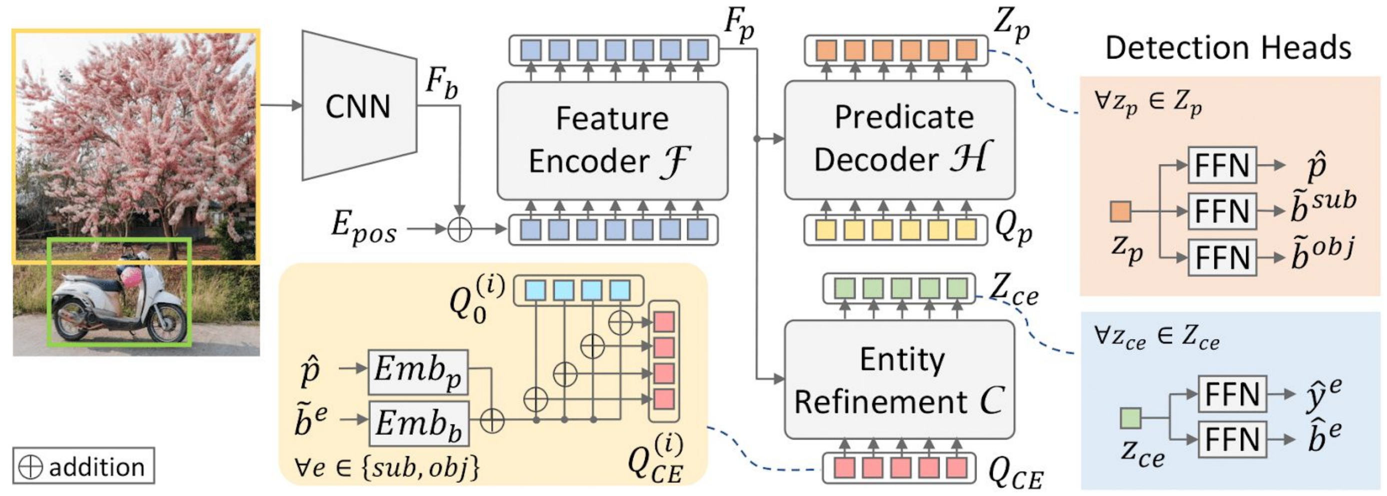
2. **Double Decoder (DD):** shared encoder and dual decoders for predicate and entity detection. Dec_p decodes <b_{sub}−p− b_{obj}> tuples and Dec_e decodes <y_{sub}− y_{obj}> pairs. Shared queries still promote entanglement.

3. **Dual DETR (DDTR):** two separate DETR models, each with an encoder, decoder and random queries. This model has the weakest entanglement between feature spaces, but is also the most expensive in terms of matching costs.

Model and Training

Transformers with conditional queries TraCQ is composed of:

- **Predicate Decoder H:** transforms a set of predicate queries into $\tilde{b}_{sub}$, $\tilde{b}_{obj}$ and $\hat{p}$
- **Entity Refinement C:** uses H's estimates to propose a set of refined bounding boxes, conditioned on the predicate label estimate



Model architecture of TraCQ

- Trained with a DETR like set prediction loss for triplet detection and Hungarian bipartite matching
- With $\hat{\sigma}_H$ as the matching of H, $\hat{\sigma}_C$ that of C and L_{cls} as the cross-entropy loss, we have

$$\mathcal{L}_p = \sum_{i=1}^{N_p}\left[\lambda_{lbl}\mathcal{L}_{cls}(\hat{p}_j, p_i) + 1_{\{p_i\neq\phi\}}\left\{\mathcal{L}_{box}(\tilde{b}_j^{sub}, b_i^{sub}) + \mathcal{L}_{box}(\tilde{b}_j^{obj}, b_i^{obj})\right\}\right]\Big|_{j=\hat{\sigma}_{\mathcal{H}}(i)}$$

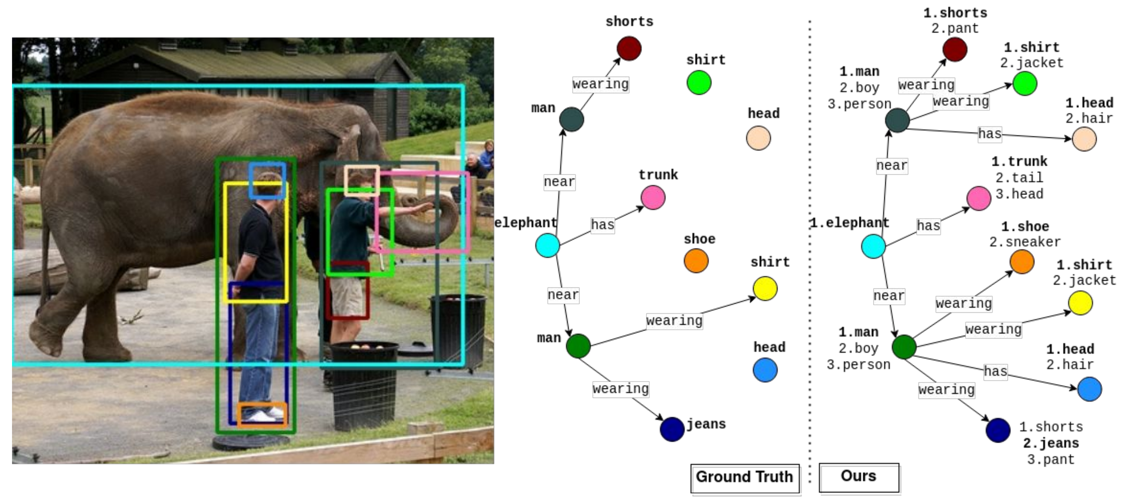
$$\mathcal{L}_e = \sum_{i=1}^{N_p} 1_{\{p_i\neq\phi\}}\left\{\sum_{j=1}^{N_{ce}}\left[\lambda_{lbl}\mathcal{L}_{cls}(\hat{y}_k^e, y_j^e) + 1_{\{y_j^e\neq\phi\}}\mathcal{L}_{box}(\hat{b}_k^e, b_j^e)\right]\Big|_{k=\hat{\sigma}_C(j)}\right\}$$

F and H are trained using L_p, whereas C is trained with L_e

Qualitative Results

Visual example generated by TraCQ shows that it can generate meaningful descriptions of the scene

1. TracQ predicts relation "wearing", between ***man*** and ***shoe***, missing in the ground truth
2. Entity labels predicted by TraCQ are synonyms of ground-truth, ***shorts/pants*** and ***shoe/sneaker***



Quantitative Results

	Method	mean-Recall (↑)			Recall (↑)			#Params (↓)
		@20	@50	@100	@20	@50	@100	
Two-Stage	MOTIFS	4.2	5.7	6.6	21.4	27.2	30.5	30.5
	BGNN	7.5	10.7	13.6	**23.3**	**31.0**	34.6	341.9
	VCTree-TDE	6.3	9.3	11.1	14.3	19.6	23.2	360.8
One-Stage	RelTR	5.8	8.5	-	20.2	25.2	-	63.7
	Relationformer	4.6	9.3	10.7	22.2	28.4	31.3	92.9
	TraCQ (ours)	**12.0**	**13.8**	**14.6**	19.7	28.3	**35.7**	**51.2**

Ablation Studies

We validate the effectiveness of the proposed formulation by ablating and observing a drop in performance going from

- Predicate first to entity first
- Predicate conditioned queries to random queries

Ablations on order of detection

Model	mean-Recall (↑)		
	@20	@50	@100
Entity-first	11.2	12.3	12.7
Predicate-first (ours)	12.0	13.8	14.6

Ablation on conditioned queries

Conditioned queries	mean-Recall (↑)		
	@20	@50	@100
w/o $Emb_p(\hat{p})$	10.8	12.4	13.1
Q_{ce} (ours)	12.0	13.8	14.6

Conclusion

- In SGG the entity and predicate spaces are entangled yet distinct
- Predicate first paradigm allows for lesser entanglement between the spaces and better visual learning compared to entity first approach
- Conditional queries allow for a smaller model and efficient inference
- TraCQ significantly outperforms existing single-stage methods, and several state-of-the-art two-stage methods as well

Project website