Tensorize, Factorize and Regularize: Robust Visual Relationship Learning

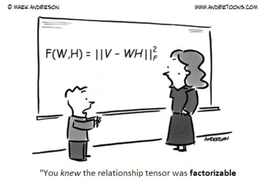
Seong Jae Hwang Sathya N. Ravi Zirui Tao Hyunwoo J. Kim Maxwell D. Collins Vikas Singh
http://pages.cs.wisc.edu/~sjh

CVPR 2018 SALT LAKE CITY • JUNE 18-22

THE UNIVERSITY of WISCONSIN MADISON

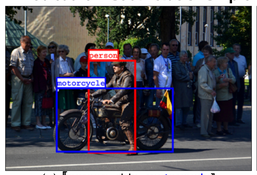
Objectives

Regularize the Scene Graph learning deep network via multi-relational tensor factorization for robust visual relationship learning.



Problem: Visual Relationship Detection

- **Given**:
  1. Image with **Object** i **and Object** j of interest
  2. **Bounding boxes** and **features** (i.e., Faster R-CNN)
  3. **Object labels** (ground truths or detection results)
- **Goal**:
  Predict the **visual relationship** of the objects: ⟦ **Object** i, **?**, **Object** j ⟧


(a) ⟦person, ride, motorcycle⟧


(b) ⟦person, ?, horse⟧

(c) Objects interactions with predicates

Figure: (a): A relationship instance in a training set. (b): An unknown relationship to predict. (c): The interactions of the objects (i.e., motorcycle and horse are both 'ridable') can be used to infer the correct relationship.

Challenges in Semantic Inference

- **Large observation space**:
  - ▷ N object categories & M possible predicates
    ⇒ N^2M possible combinations
- **Sparse observations**:
  - ▷ Visual Genome has 1M relationship instances
    ⇒ But observed only ∼2% of possible combinations
- **Zero-shot learning**:
  - ▷ Inferring cases unobserved in train set
    ⇒ ∼98% of possible cases *NOT* observed in train set

Zero-shot Learning: Unobserved Visual Relationship Detection


⟦horse, wear, hat⟧


⟦mouse, on, cabinet⟧

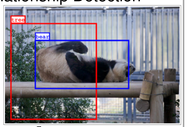

⟦tree, behind, bear⟧

Figure: The *unobserved* observed relationships are potentially much harder to detect.

Step 1: Tensorize the Visual Relationships

- **Multi-relational tensor** $X \in \mathbb{R}^{n\times n\times m}$ given n object categories and m possible predicates
- $X(i,j,k)$: number of ⟦Object i, Predicate k, Object j⟧ in train set
- ∼2% is non-zero ⇒ **extremely sparse tensor**

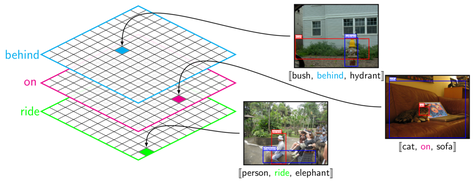


Figure: Each 'slice' X_k encodes possible relationships involving its corresponding k^{th} predicate.

Step 2: Factorize the Relational Tensor

Multi-relational Tensor Factorization

- Based on the multi-relational tensor X from train set, derive
  1. **Common** latent representation of objects $A \in \mathbb{R}^{n\times r}$
  2. **Relationship-specific** factor matrix $R_k \in \mathbb{R}^{r\times r}$ for each relationship $k \in \{1,\dots,m\}$

⇒ such that $X_k \approx AR_kA^T$

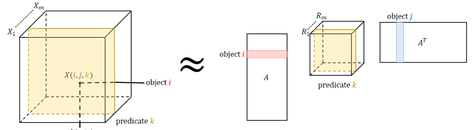


Figure: Multi-relational tensor factorization $X_k \approx AR_kA^T$ for $k \in \{1,\dots,m\}$.

Multi-relational Tensor Factorization

$$\min_{A,R_k} \sum_{k=1}^{m} ||X_k - AR_kA^T||_F^2 \quad (1)$$

1. **4^{th}-order term** A: Use auxiliary variables to decouple A and A^T:
$$\min_{A,R_k} \sum_{k=1}^{m} ||X_k - B_kA^T||_F^2 \quad \text{s.t.} \quad B_k = AR_k. \quad (2)$$
2. **Low-rank Initialization via SVD**: For $m = 1$, $U\Sigma V^T = X$. Initialize with the "**basin of attraction**" (Luo et al., Tu et al.):
$$A = V\Sigma^{1/2}, \quad B = U\Sigma^{1/2}.$$
3. **Restrict degenerate cases**: $A' = AP^{-T}$ and $B' = BP^{-T}$ for any invertible P. Normalize A and B_k:
$$\lambda_p \sum_{k=1}^{m} ||A^TA - B_k^TB_k||_F^2. \quad (3)$$

Final Formulation

$$\min_{A,R_k,B_k} \sum_{k=1}^{m} ||X_k - B_kA^T||_F^2 + \gamma \sum_{k=1}^{m} ||B_k - AR_k||_F^2 + \lambda_p \sum_{k=1}^{m} ||A^TA - B_k^TB_k||_F^2 \quad (4)$$

Algorithm: Alternating Block Coordinate Descent

Algorithm 1 Alternating Block Coordinate Descent on (4)

Given: $X \in \mathbb{R}^{n\times n\times m}$, $X_k := X(:,:,k)$, rank $r > 0$
Low-rank Initialization:
$\overline{X} \leftarrow \sum_{k=1}^{m} X_k$
$\overline{U}\overline{\Sigma}\overline{V}^T \leftarrow \text{SVD}(\overline{X}, r)$
$A \leftarrow \overline{V}\overline{\Sigma}^{1/2}$
for $k = 1, \dots, m$ **do**
  $B_k \leftarrow \overline{U}\overline{\Sigma}^{1/2}$
  $R_k \leftarrow (A^TA)^{-1}(A^TX_kA)(A^TA)^{-1}$
end for
Iterative descent method:
while Convergence criteria not met **do**
  $A \leftarrow$ gradient descent on (4) w.r.t. A
  for $k = 1, \dots, m$ **do**
    $B_k \leftarrow$ gradient descent on (4) w.r.t. B
    $R_k \leftarrow (A^TA)^{-1}(A^TB_k)$
  end for
end while
Output: $A \in \mathbb{R}^{n\times r}$, $B_k \in \mathbb{R}^{n\times r}$, $R_k \in \mathbb{R}^{r\times r}$ for $\forall k$

Experiment 1: Predicate Detection

Scene Graph dataset:

- 5000 training (< 1% unique relationships), 1000 test images
- $n = 100$ object categories, $m = 70$ predicates
- **Given**: Test images with FastRNN bboxes / labels
- **Goal**: Predict ⟦ Object, Predicate, Object⟧ with 3 tasks

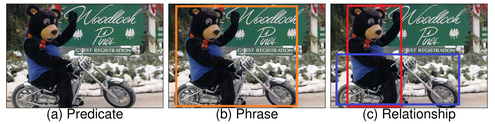
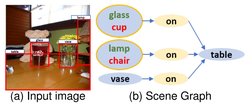

(a) Predicate (b) Phrase (c) Relationship

Figure: **Detection Task Conditions:** (a) **Predicate** (easy): does not require bounding boxes. (b) **Phrase** (moderate): requires relationship bounding box (orange) containing both objects. (c) **Relationship** (hard): requires individual bounding boxes (red/blue).

Experiment 2: Scene Graph Detection

Visual Genome dataset:

- 108,077 rels. (< 2% unique relationships), 70% training, 30% test
- $n = 150$ object categories, $m = 50$ predicates
- **Given**: Test images with Region Proposal Network bboxes / labels
- **Goal**: Predict Scene Graphs with 3 standards


(a) Input image (b) Scene Graph

Prediction Tasks	Pred.	Obj.	B-box
Predict Predicate (**PredCls**)	✓		
Classify SG (**SgCls**)	✓	✓	
Generate SG (**SgGen**)	✓	✓	✓

Figure: Scene graph detection tasks. Check marks indicate required prediction components. The tasks become incrementally more demanding from top (PredCls) to bottom (SgGen).

Experiment 1: Results

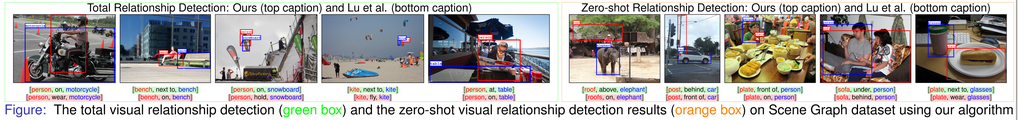


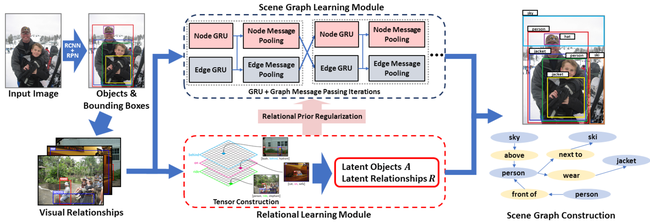
Figure: The total visual relationship detection (green box) and the zero-shot visual relationship detection results (orange box) on Scene Graph dataset using our algorithm (top caption) and Lu et al. (bottom caption). The correct and incorrect predictions are highlighted in green and red respectively.

Figure: **Left**: Examples of semantically confusing relationships of the ground truth (top caption), ours (middle caption) and Lu et al. (bottom caption). The left three relationships (green box) have varying detection results, but they can all be considered to be semantically correct. However, the right two relationships (orange box) have disagreeing interpretations of the relationships. **Right**: Visual relationship detection results on Scene Graph.

Step 3: Regularize the Scene Graph Learning Pipeline

Scene Graph detection: Predicate detection ⇔ Object detection

- Module 1: **Scene Graph Module** (Xu et al.) ⇒ Faster-RCNN → GRU → Message Pooling → Exchange
  - ▷ *Expressive* but highly dataset dependent (i.e., prone to outliers and mislabels)
- Module 2: **Relational Learning Module** (Ours) ⇒ Faster-RCNN → Multi-relational Tensor Factorization
  - ▷ *Robust* but does not infer objects by construction
  - ▷ **Pre-trained and fixed during training**



Balance the learning modules

- Bring the best of both worlds!
- Regularize with **global prior** from RL module

Probabilistic Global Prior Injection

Given (i) SG module prediction k^*_{SG}, (ii) RL module prediction k^*_{RL}, and (iii) a 'y-or-1' filter $D(k^*_{RL}(i),\theta) = k^*_{RL}(i)$ with probability θ, we obtain the **regularized prediction** k^* as follows:

$$k^* = k^*_{SG} \odot D(k^*_{RL}, \theta) \quad (5)$$

where $\odot$ is a Hadamard product.

Result: *Increase underestimated scores* (i.e., rare relationships) or *decrease overestimated scores.* High θ, more frequent influence of prior k^*_{RL}.

Training the pipeline:

1. '**Warm start**' with *only* SG for 100K iters ($\theta = 0$)
2. **Regularize** k^*_{SG} with k^*_{RL} to obtain k^* as 5 with $\theta = 0.2$ for 50K iters (RL module is fixed)

Experiment 2: Results

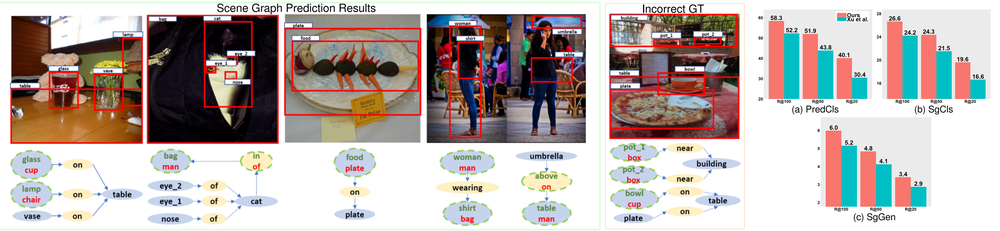


Figure: **Left**: For each column, the predicted objects (blue ellipses) and their relationships (yellow ellipses) are constructed as a scene graph of its top image. The bounding boxes labels reflect our prediction results. For difficult predictions (green dashed boundary) where our model has correctly predicted (top green) and while Xu et al. has misclassified (bottom red) are shown. The rightmost column is an example of a case where our model provides more accurate predictions (pot and bowl) than those of the ground truth (box and cup). **Right**: Scene graph detection task results on Visual Genome.

Acknowledgment

SJH was supported by a University of Wisconsin CIBM fellowship (5T15LM007359-14). We acknowledge support from NIH R01 AG040396 (VS), NSF CCF 1320755 (VS), NSF CAREER award 1252725 (VS), UW ADRC AG033514, UW ICTR 1UL1RR025011, Waisman Core grant P30 HD003352-45 and UW CPCP AI117924 (SNR).