Visual Relationship Detection with Low Rank Non-Negative Tensor Decomposition

Mohammed Haroon Dupty,* **Zhen Zhang, Wee Sun Lee**
School of Computing, National University of Singapore
{dmharoon, leews}@comp.nus.edu.sg, zhen@zzhang.org

Abstract

We address the problem of Visual Relationship Detection (VRD) which aims to describe the relationships between pairs of objects in the form of triplets of (*subject, predicate, object*). We observe that given a pair of bounding box proposals, objects often participate in multiple relations implying the distribution of triplets is multimodal. We leverage the strong correlations within triplets to learn the joint distribution of triplet variables conditioned on the image and the bounding box proposals, doing away with the hitherto used independent distribution of triplets. To make learning the triplet joint distribution feasible, we introduce a novel technique of learning conditional triplet distributions in the form of their normalized low rank non-negative tensor decompositions. Normalized tensor decompositions take form of mixture distributions of discrete variables and thus are able to capture multimodality. This allows us to efficiently learn higher order discrete multimodal distributions and at the same time keep the parameter size manageable. We further model the probability of selecting an object proposal pair and include a relation triplet prior in our model. We show that each part of the model improves performance and the combination outperforms state-of-the-art score on the Visual Genome (VG) and Visual Relationship Detection (VRD) datasets.

Introduction

Object detection is a central problem in computer vision. Recent deep learning approaches (Ren et al. 2015; Girshick et al. 2014) have made long strides in the task of object detection. However, real world images often involve multiple objects that interact with each other. Much can be said about the image if we can reason object interactions with each other in addition to detection. Reasoning about the relationships objects participate in provides a powerful method for capturing mid-level information that is useful for computer vision tasks; for example, in image captioning, richer captions can be generated if objects as well as the relationships between objects in the image is provided. The task of Visual Relationship Detection aims to recognize and localize the objects along with predicting the relationship that pairs of objects participate in.

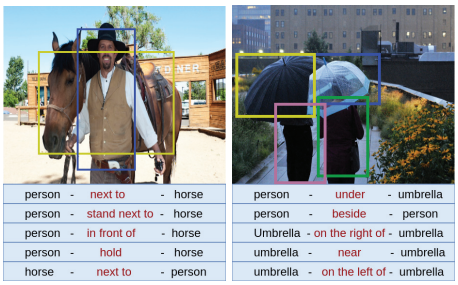

Figure 1: Visual relationships are defined with triplets of (subject - predicate - object). Multiple predicates may exist between a pair of box proposals suggesting the distribution of triplets given box proposals is often multimodal. We represent this conditional triplet distribution as normalized low rank tensor decomposition. This representation takes form of mixture distribution of triplet variables thus capturing the multimodality while being efficiently learnable.

State-of-the-art methods for visual relationship detection (Dai, Zhang, and Lin 2017; Xu et al. 2017; Liang, Lee, and Xing 2017; Zhu and Jiang 2018) solve the problem in two stages: an objects proposal stage that uses object detection to propose a set of object bounding box proposals that may participate in relations within the image, and a relation recognition stage that outputs a set of possible relation triplets given the set of bounding box proposals provided by the objects proposal stage. In this paper, we focus on the relation recognition stage. Our main observation is that multiple relations often exist between a pair of box proposals in an image: for example, in Figure 1, the relations <*person - next to - horse*>, <*person - in front of - horse*> and <*person - hold - horse*> are all valid. This suggests that the distribution of triplets, given a pair of objects, is often multimodal with multiple valid relations occurring between a pair of objects. Any learning model with a single separate output for object and predicate classes cannot represent such a multimodal distribution.

In this work, we use a neural network to model the prob-

*corresponding author

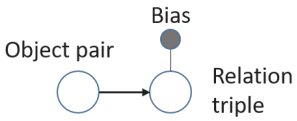

Figure 2: Graphical model representing the relation recognition model. Given an input image and its bounding boxes, a pair of bounding boxes is selected by the annotator and then annotated with a relation triplet. The relation triplet node is augmented by a single node potential representing the bias caused by the prevalence of each triplet in the dataset.

ability of a triplet $p(t|b, I)$ where $t = (x_s, x_p, x_o)$ consists of the subject, predicate and object labels, $b = (b_s, b_o)$ is the pair of box proposals and I is the image. However, providing a separate output for each triplet combination requires a large number of outputs and a correspondingly large number of parameters in the network; learning such a network would require a large training set. Instead of providing a separate output for each triplet, our neural network outputs a low-rank non-negative tensor decomposition, given each image and pair proposal as input. This limits the number of outputs to be proportional to the sum (instead of product) of the number of values that each variable in the triplet can take. Further, tensor decomposition structure enables computing gradient of loss without construction of full tensor which makes training efficient.

The frequency of appearance of the triplets, independent of the proposed bounding box pair, provide a useful bias for improving performance. To incorporate information about the unconditional frequency of appearance of each triplet, we multiply each conditional output with a bias where the bias is represented using a three dimensional non-negative tensor of triplet frequencies. This gives a simple conditional random field representation for the conditional distribution of triplets $p(t|b) = \psi_c(t|b)\psi_u(t)$ where ψ_c is the conditional potential function represented using a neural network that outputs low-rank non-negative decomposition, and ψ_u is the unconditional potential function to capture the frequency bias, represented using a three dimensional tensor.

In addition to multi-modality, previous studies (Lu et al. 2016; Krishna et al. 2017) have shown that there exists missing annotation problem in all visual relationship datasets, i.e. a relationship is annotated in certain examples and the same relationship is not annotated in other examples though it exists; for example, it is more likely that pairs of boxes close to each other catch the attention of the annotator than pairs which are farther apart. We model that with probability of a pair of boxes being selected by the annotator for annotation.

The complete process of generating the relation triplets given an image and a set of detected boxes can be represented with a probabilistic model shown in Figure 2. In summary, we make following contributions:

- We propose a novel way of learning higher order triplet distributions in the form of their low rank tensor decompositions. This representation takes form of the mixture of rank-1 tensors and thereby is able to represent multimodal distributions.
- Our formulation enables efficient computation of gradient of the log likelihood and errors can be backpropagated without forming the higher order tensor.
- We further augment the conditional triplet distribution with relation frequency prior and probability of annotating a object proposal pair which together outperform the state-of-the-art score on visual relationship detection.

Related work

We need compact form representation of joint distribution of triplets and tensor decompositions are a natural alternative to represent such higher-order functions. There has been a recent surge in using tensor decompositions in various machine learning problems (Anandkumar et al. 2014), (Janzamin, Sedghi, and Anandkumar 2015), (Wrigley, Lee, and Ye 2017). Low rank representation has been shown to perform reasonably well particularly when the size of the exact model is large. For the task of visual relationship detection in (Jae Hwang et al. 2018), the empirical distribution of visual relationships in the dataset was approximated with low rank Tucker decomposition and used as a prior for regularization during learning. In contrast to this work, we use tensor decomposition to represent the conditional probability distribution of the relation triplets.

Over the last few years, a number of different approaches have been proposed for the task of recognizing relationships from the image. Most of these methods can be divided into three broad categories. One line of work uses structured prediction techniques by message passing among the three triplet variables (Dai, Zhang, and Lin 2017; Xu et al. 2017; Liang, Lee, and Xing 2017; Zhu and Jiang 2018). Structured prediction techniques are mainly useful as the predicate distribution conditioned on the object labels is highly predictive as was shown in (Dai, Zhang, and Lin 2017). They take into account within triplet dependencies by message passing among object and predicate labels.

Another line of work introduces extra information either in the form of word vector embeddings of the object labels or use knowledge from a large corpus (Lu et al. 2016; Yu et al. 2017; Zhang et al. 2017a; 2019a). Learning from a large external knowledge has been shown to be an effective strategy to tackle the missing data problem (Yu et al. 2017). Other prevalent methods (Lu et al. 2016; Zhang et al. 2019a; 2017a; Yu et al. 2017) have used a combination of visual and linguistic features to detect relationships between objects and showed the utility of word vector embeddings. With much work done, it is now well established that the use of word vectors is substantially helpful in recognizing relationships (Lu et al. 2016; Yu et al. 2017; Zhang et al. 2017a; 2019a).

A third line of work uses rank-based loss functions to encourage similar relations to be close to each other in the learnt feature space (Liang et al. 2018; Zhang et al. 2019a). Most recently (Zhang et al. 2019a) used triplet loss to match the visual and semantic features in a projected shared space for better discriminative power.

In this work we look only from the visual perspective without any sort of external information. Instead, we try to capture the multimodal properties of the triplet distribution and to model the generative annotation process.

Preliminaries: Tensor decomposition

Tensors are generalizations of matrices to higher dimensions and hence a tensor can be called a multidimensional array. A order-d tensor T is an element in $\mathcal{R}^{N_1 \times N_2 \cdots \times N_d}$ with N_k possible values in k^{th} dimension. Analogous to SVD in matrices, tensors can be represented in succinct form with tensor decompositions. In CANDECOMP / PARAFAC (CP) decomposition, any tensor T can be represented as a linear combination of outer products of vectors as

$$T = \sum_{r=1}^{R} w_r \phi_{r,1} \otimes \phi_{r,2} \otimes \cdots \otimes \phi_{r,d} \tag{1}$$

where $\otimes$ is the outer product operator, each $\phi_{r,k}$ is a vector in R^{N_k} for $k \in \{1, 2, \ldots, d\}$ and the term $\phi_{r,1} \otimes \phi_{r,2} \otimes \cdots \otimes \phi_{r,d}$ is a rank-1 tensor. w_r is a scalar co-efficient which can be absorbed in one of the vectors $\phi_{r,k}$. Tensor value at index $(i_1, i_2 \ldots, i_d)$ is given by $\sum_{r=1}^{R} w_r \phi_{r,1}^{i_1} \phi_{r,2}^{i_2} \cdots \phi_{r,d}^{i_d}$. The smallest R for which an exact R-term decomposition exists is the rank of tensor T and the decomposition (1) is its R-rank approximation of the tensor. With this compact representation a tensor T with $N_1 \times N_2 \cdots \times N_d$ entries can be represented with R vectors for each variable in T i.e. with $R(N_1 + N_2 + \cdots + N_d)$ entries. More information about tensor decompositions can be found in (Kolda and Bader 2009; Rabanser, Shchur, and Günnemann 2017).

With low rank assumption, we can represent any probability distribution of multiple variables in tensor form of (1), provided we constrain the values of T to be non-negative and normalize it such that sum of all entries of T is 1. We call such a normalized non-negative tensor a ***probability tensor***. Note that the form of the ***probability tensor*** in (1) is more like mixture distribution of discrete variables and thus is suitable for modeling multimodal triplet distribution.

Formulation

Like other recent works (Dai, Zhang, and Lin 2017; Liang, Lee, and Xing 2017; Zhu and Jiang 2018), we treat task of visual relationship detection with a two stage pipeline where the boxes are given by a separately trained object detector, Faster-RCNN (Ren et al. 2015) and a classifier predicts the relationships between each pair of the boxes including a null relation for box pair that do not participate in a relation.

Formally given an image I with N objects represented by rectangular bounding box proposals, we have $N(N-1)$ relations between each pair of the objects. For each pair of subject and object box proposal $(B_s = b_s, B_o = b_o)$, subject label $X_s = x_s$, predicate label $X_p = x_p$ and object label $X_o = x_o$ form a triplet instance.

Conditional triplet distribution

We use two sources of information for modeling the conditional triplet distribution $P(X_s, X_p, X_o | B_s, B_o, I)$. The first source is information available in the image and bounding box pair. The second is the prior distribution of triplets, not conditioned on the image. We represent the conditional triplet distribution as a product of the two potential functions $P(X_s, X_p, X_o | B_s, B_o, I) = \psi_c(X_s, X_p, X_o | B_s, B_o, I) \cdot \psi_u(X_s, X_p, X_o)$ to give a simple conditional random field.

$\psi_u(X_s, X_p, X_o)$ is constructed from the training set. It is a order-3 tensor representing the number of times each triplet occurs in the dataset. It serves as a frequency bias towards most frequently occurring relations. $\psi_u(X_s, X_p, X_o) \in \mathbb{R}^{|X_o| \times |X_p| \times |X_o|}$ (for $|X_o|$ object and $|X_p|$ predicate classes). The value at $\psi_u(i, j, k)$ is the number of times triplet (x_s^i, x_p^j, x_o^k) occurs in the training set where x_s^i is the i^{th} subject class, x_p^j is the j^{th} predicate class and x_o^k is the k^{th} object class. It turns out that, only $\sim 1\%$ of the tensor constructed from the training set has non-zero entries. Multiplication with ψ_c will cause all unseen relation triplets to vanish. To address this issue, we smoothen ψ_u by adding 1 to all entries. We then normalize ψ_u by dividing with its sum to get prior probability of occurrence of all relation triplets.

$\psi_c(X_s, X_p, X_o | B_s, B_o, I) \in \mathbb{R}^{|X_o| \times |X_p| \times |X_o|}$ is also an order-3 tensor i.e. given an image and a pair of boxes, ψ_c gives probability of triplet labels. We assume that it is well approximated with a low rank tensor. Our assumption stems from the observation that out of all possible relationships, only certain specific relationships tend to occur frequently given the underlying objects. The sparsity of prior ψ_u further strengthens our low rank assumption of the tensor. As the rank of any tensor cannot exceed the number of non-zero entries in the tensor, assuming that the tensor is low rank appears to be reasonable. Consequently, we use a mixture of independent rank-1 tensors to represent $\psi_c(X_s, X_p, X_o | B_s, B_o, I)$, allowing us to effectively capture richer multimodal triplet distributions with a reasonably compact model. We represent the order-3 tensor $\psi_c(X_s, X_p, X_o | B_s, B_o, I)$, in CP-decomposition form as

$$\psi_c(X_s, X_p, X_o | B_s, B_o, I) = \sum_{r=1}^{R} w_r \phi_{rs}(X_s | B_s, B_o, I) \otimes \phi_{rp}(X_p | B_s, B_o, I) \otimes \phi_{ro}(X_o | B_s, B_o, I)). \tag{2}$$

For notational convenience, we omit the conditioning on the image and bounding boxes in ϕ_{ra} from here onwards, where $\phi_{ra}(X_a)$ is the r^{th} vector of the variable X_a for each $a \in \{s, p, o\}$. We parameterize $\psi_c(X_s, X_p, X_o | B_s, B_o, I)$ with a deep neural network and learn it from the data. Given an image and a pair of bounding boxes, the neural network outputs a set of R vectors $s_{a,r}$ for $a \in \{s, p, o\}$. Since potential functions are required to be non-negative, we represent $\phi_{ra}(i) = e^{s_{a,r}^i}$ so that the tensor decomposition is non-negative. Then, we normalize the output tensor by dividing with the sum of all tensor entries to make it a ***probability tensor***.

Without loss of generality, we assume the weights w_r in eqn (2) can be absorbed into the vectors ϕ_{ra} and hence there no need for their separate representation. Our representation reduces the number of outputs required to represent the

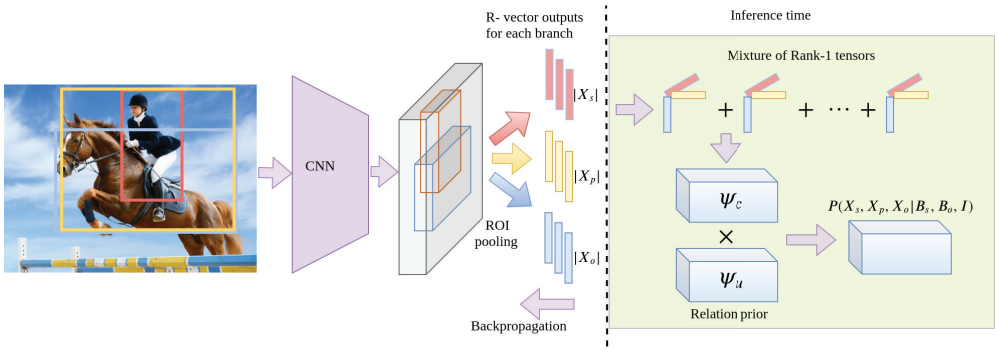

Figure 3: An image is input to a neural network (VGG16) to produce an intermediate feature map. For each pair of boxes from the detector, the corresponding features are ROI-pooled and fed through three separate branches of fully connected layers each for subject, predicate and object. Each branch outputs a set of R vectors which together form the mixture of independent triplet distributions(ψ_c) capturing the multimodal distribution. ψ_c is multiplied with unconditional relation prior ψ_u constructed from the training set to give $P(X_s, X_p, X_o|B_s, B_o, I)$. During training, construction of ψ_c is not required and errors can be backpropagated from the set of R vectors.

function from $|X_s|\times|X_p|\times|X_o|$ to $R(|X_s|+|X_p|+|X_o|)$. A smaller sized output corresponds to a smaller number of model parameters, making it possible to learn the model from less data. Finally, prior $\psi_u(X_s, X_p, X_o)$ is multiplied elementwise with $\psi_c(X_s, X_p, X_o|B_s, B_o, I)$ to get the conditional triplet distribution $P(X_s, X_p, X_o|B_s, B_o, I) = \psi_c(X_s, X_p, X_o|B_s, B_o, I)\psi_u(X_s, X_p, X_o)$. Note that due to low rank assumption of the ψ_c, the network may output spurious triplets which may not be seen in the training set. Probablities of such spurious triplets are pushed down by multiplication with the prior.

Training Loss

We learn the triplet distribution $P(X_s, X_p, X_o|B_s, B_o, I) = \psi_c(X_s, X_p, X_o|B_s, B_o, I)\ \psi_u(X_s, X_p, X_o)$ using a deep neural network. As both ψ_c and ψ_u are normalized to sum to 1, they both can be trained separately by treating $\psi_c(X_s, X_p, X_o|B_s, B_o, I)$ as a conditional distribution and $\psi_u(X_s, X_p, X_o)$ as a prior distribution. This type of learning is often called piecewise training of the conditional random field (Lin et al. 2016). The prior distribution is learned simply by computing the frequencies of the triplets in the training set.

For conditional distribution $\psi_c(X_s, X_p, X_o|B_s, B_o, I)$, we use a neural network that outputs a set of vectors that correspond to the tensor decomposition, given an image and pair of bounding boxes. The network outputs a set of R vectors each for subject, predicate and object for a total of $3R$ vectors. Each of these R vectors are indexed by i, j, k for subject, predicate and object categories respectively. Let $s_{a,r}$ be the final layer output of the network for each $a \in \{s, p, o\}$. We exponentiate $s_{a,r}$ and then normalize to make it a ***probability tensor*** y. $(i, j, k)^{th}$ entry of y can be computed as

$$y^{i,j,k} = \frac{\sum_{r=1}^{R} e^{s^i_{s,r}} e^{s^j_{p,r}} e^{s^k_{o,r}}}{\sum_{l,m,n}\sum_{r=1}^{R} e^{s^i_{s,r}} e^{s^j_{p,r}} e^{s^k_{o,r}}}. \tag{3}$$

For training such a network, we use cross-entropy loss as it is a classification problem. But unlike the usual case, the loss is computed between tensors instead of vectors. If t representss the target one-hot tensor (tensor that is zero everywhere, except at a single tuple index which has value 1, representing the indicator function of the tuple) and y output *probability tensor*, the cross-entropy loss function is

$$L = -\sum_{i,j,k} t^{i,j,k} \log(y^{i,j,k}). \tag{4}$$

A simple brute force method to compute eqn (4) is to fully construct tensor y from the network output and then compute loss. But this would significantly slow down training. Instead we compute the derivative of loss with respect to final layer output s_a directly without constructing probability tensor y.

Consider the derivative of the loss w.r.t $s^i_{s,r'}$ where subscript s indicates subject variable X_s, r' is one of the R vectors and i is the i^{th} index out of $|X_s|$ indices. For an observed tuple (i', j', k'), the derivative of the Loss L with respect to $s^i_{s,r'}$ is given by:

$$\frac{\partial L}{\partial s^i_{s,r'}} = \begin{cases} \dfrac{e^{s^i_{s,r'}} \sum_{m'} e^{s^{m'}_{p,r'}} \sum_{n'} e^{s^{n'}_{o,r'}}}{Z} & i \neq i' \\ \dfrac{e^{s^i_{s,r'}} \sum_{m'} e^{s^{m'}_{p,r'}} \sum_{n'} e^{s^{n'}_{o,r'}}}{Z} & \\ -\dfrac{e^{s^i_{s,r'}+s^{j'}_{p,r'}+s^{k'}_{o,r'}}}{\sum_{r=1}^{R} e^{s^i_{s,r}+s^{j'}_{p,r}+s^{k'}_{o,r}}} & i = i' \end{cases}$$

where

$$Z = \sum_{l,m,n}\sum_{r=1}^{R} e^{s^l_{s,r}+s^m_{p,r}+s^n_{o,r}} \tag{5}$$

is the partition function.

Naive computation of Z by summing over all entries of tensor y may significantly slow down training. Instead we can compute the partition function efficiently in time linear with class size of each variable X, by simply pushing the

outer sum $\sum_{l,m,n}$ inside and only evaluate it over the corresponding univariate vectors i.e.

$$Z = \sum_{r=1}^{R} \sum_{l} e^{s^{l}_{s,r}} \sum_{m} e^{s^{m}_{p,r}} \sum_{n} e^{s^{n}_{o,r}}. \quad (6)$$

The derivatives with respect to the predicate and object variable outputs $e^{s_{p,r'}}$ and $e^{s_{o,r'}}$ are computed in a similar way and backpropagated to optimize the network.

Modeling missing annotations

The problem of missing annotations in relationship detection datasets is well known (Lu et al. 2016; Krishna et al. 2017). A relationship is annotated in certain examples and the same relationship may not be annotated in other examples though it exists. We have $N(N-1)$ pairs for N object boxes and only few of them are annotated with relations, rest of the pairs are considered *null*. Some of these *null* pairs may have valid relations but are not annotated. This confuses the model during training as examples with similar features are considered valid as well as *null*. We handle this problem by training $P(X_s, X_p, X_o|B_s, B_o, I)$ with only annotated positive relations. We train a separate binary variable with equal number valid and *null* examples to give $P(X_{sel}|B_s, B_o, I)$, probability of annotation of box pair. At test time, $P(X_{sel}|B_s, B_o, I))$ is multiplied with $P(X_s, X_p, X_o|B_s, B_o, I))$ before final ranking. With this technique, triplet network produces reliable values for valid relations and the unreliable values produced for box-pairs with *null* relation are brought down by multiplication with $P(X_{sel}|B_s, B_o, I)$.

The final scoring function f_{spo} for each of the tuple $(X_s, X_p, X_o, B_s, B_o)$ is given by the full posterior of the relation triplet and pairs of bounding boxes for a given image.

$$P(B_s, B_o|I)P(X_s, X_p, X_o|B_s, B_o, I)P(X_{sel}|B_s, B_o, I) \quad (7)$$

where $P(B_s, B_o|I)$ is from detector output.

Network Architecture

Figure (3) shows the workflow of our model. We use VGG16 (Simonyan and Zisserman 2014) as backbone of our network initialized with pretrained weights on Visual Genome for detection with freezed conv1_1-conv5_3 layers and train it with the ground truth gold proposal boxes. We feed the network with the image to get a global feature map of the image from which subject, predicate and object features are ROI-Align pooled w.r.t their corresponding box regions. The predicate feature is pooled from the union-box of the subject and object boxes. After ROI-pooling, visual features are fed through three separate branches each for subject, object and predicate.

Parallelly, we include 2-channel spatial binary mask feature of size $2 \times 64 \times 64$ as in (Dai, Zhang, and Lin 2017). Each channel is a matrix with 1 in bounding box region of object (scaled to size 64×64) and 0 everywhere. This feature is passed through 2 convolution layers and then a fully connected layer to get a 512-dim *spatial feature* which is concatenated with ROI-pooled predicate feature. Each of the subject, predicate and object branches has two fully connected layers of size 4096 with final layer output size of $R \times |X_a|$ for $a \in \{s, p, o\}$.

This ROI-pooled predicate feature concatenated with *spatial feature* also serves as the input for binary classifier for X_{sel} with a single hidden layer of 4096. We first train the triplet network with only annotated positive relationships. We then freeze weights of triplet network and train X_{sel} with equal number of positive and negative examples.

Implementation details: We implement our method on pytorch, a mainstream deep learning library. We set the learning rate to 1e-4 and use SGD as optimizer. Due to the summation in the gradient term, there is an exploding gradient problem. To fix this, we clip the gradient based on the total norm of all the learnable weights. The norm value for gradient clipping is set at 20. We then train with proposals from the detector. We sample atmost 4 proposals for every ground truth box proposal with IOU overlap of atleast 0.5. All the layers before ROI-pooling are initialized by pretrained weights from the detector. During inference, we filter out the overlapping boxes from the detector by enforcing Non-Maximum Suppression (NMS) constraints with NMS threshold set at 0.7[1].

Computation time: With our low-rank tensor formulation of ψ_c, we are able to train the triplet network with VRD dataset at 6 min/epoch and VG dataset at 90min/epoch for 7 epochs. Backpropagation on a single image takes 0.21sec on average. Inference for a single image takes around 0.57s with gold proposals. Training time per image is substantially low compared to its inference time as we do not construct full 3D tensor ψ_c during training.

Experiments

We evaluate our method on the Visual Genome and the Visual Relationship Detection datasets.

VG: The Visual Genome dataset was released by (Krishna et al. 2017). Unfortunately there is not a single version of Visual Genome that is consistently used by all previous works on this task. To show better performance of our model across splits of VG, we use two different versions in our experiments, **VG200** (Zhang et al. 2017a) and **VG150** (Xu et al. 2017) for comparison with other recent works. **VG200** has 200 object and 100 predicate categories. **VG150** has 150 object and 50 predicate categories. We conduct our ablation studies on VG150.

VRD: The VRD dataset was released by (Lu et al. 2016) with standard train/test split 4000 and 1000 images respectively. There are 100 object and 70 predicate categories with $6,672$ unique relationships. On average there are 24.25 relationships per object category.

Evaluation tasks: Consistent with prior works, we report results on two tasks, Relationship Detection and Phrase Detection. In both the tasks we are given an input image and required to output top-50/100 relation triplets with the corresponding bounding boxes for each pair. In **Relationship detection**, a prediction is considered correct if all three triplet labels (s,p,o) are correctly recognized, and the intersection

[1]https://github.com/dmharoon/VRD-Tensor-Decompostion

	Relationship		Phrase		Relationship detection				Phrase detection			
	mult preds (free k)				k=1		k=10		k=1		k=10	
Recall at	50	100	50	100	50	100	50	100	50	100	50	100
w/ proposals from (Lu et al. 2016)												
Language Cues (Plummer et al. 2017)	16.89	20.70	15.08	18.37	-	-	16.89	20.70	-	-	15.08	18.37
VRD (Lu et al. 2016)	17.43	22.03	20.42	25.52	13.80	14.70	17.43	22.03	16.17	17.03	20.42	25.52
LargeVRU (Zhang et al. 2019a)	19.18	22.64	21.69	25.92	16.08	17.07	19.18	22.64	18.32	19.78	21.69	25.92
Ours	**24.08**	**28.29**	**29.17**	**34.33**	**17.67**	**18.64**	**24.08**	**28.29**	**20.80**	**22.13**	**29.17**	**34.33**
w/ better proposals												
L distilation(Yu et al., 2017)	22.68	31.89	26.47	29.76	19.17	21.34	22.56	29.89	23.14	24.03	26.47	29.76
Zoom-Net (Yin et al. 2018)	21.37	27.30	29.05	37.34	18.92	21.41	-	-	24.82	28.09	-	-
CAI + SCA-M (Yin et al. 2018)	22.34	28.52	29.64	38.39	19.54	22.39	-	-	25.21	28.89	-	-
LargeVRU (Zhang et al. 2019a)	26.98	32.63	**32.90**	39.66	23.68	26.67	26.98	32.63	28.93	32.85	**32.90**	39.66
MF-URLN (Zhan et al. 2019)	23.9	26.8	31.5	36.1	23.9	**26.8**	-	-	**31.5**	**36.1**	-	-
Ours	**27.09**	**34.93**	32.29	**41.28**	**24.20**	25.87	**27.09**	**34.93**	28.53	30.92	32.29	**41.28**

Table 1: Comparison with state of the art methods on VRD dataset.

Method	Relation Detection		Phrase Detection	
	R@50	R@100	R@50	R@100
VTranseE (Zhang et al. 2017a)	5.5	6.0	9.5	10.5
PPRFCN (Zhang et al. 2017b)	6.0	6.9	10.6	11.1
DSL (Zhu and Jiang 2018)	6.8	8.0	13.1	15.6
VSA (Han et al. 2018)	6.0	6.3	9.7	10.0
MF-URLN (Zhan et al. 2019)	14.4	16.5	26.6	32.1
Ours (k=1)	**16.74**	**18.69**	**29.32**	**33.42**
Ours (free k)	**18.52**	**21.92**	**31.58**	**38.07**

Table 2: Comparison with state-of-the-art on VG200 dataset

Method	PhrCls (free k)		PhrCls (k=1)	
	R@50	R@100	R@50	R@100
Freq-Overlap (Zellers et al. 2018)	39.0	43.4	32.3	32.9
Message Passing (Xu et al. 2017)	43.4	47.2	34.6	35.4
Motifnet (Zellers et al. 2018)	44.5	47.7	35.8	36.5
RelDN (Zhang et al. 2019b)	**48.9**	50.8	**36.8**	36.8
Ours	47.54	**54.69**	35.60	**37.68**

Table 3: Comparison with state-of-the-art on VG150 dataset

over union (IOU) between the predicted and the ground-truth boxes is at least 0.5. In **Phrase detection** the prediction is correct if triplet labels match and IOU of the union of the two predicted boxes with the union of ground-truth boxes is at least 0.5. In line with previous works (Yu et al. 2017; Zhang et al. 2019a), we consider k relationship predictions per object box pair before taking the top-50/100 predictions per image. We report for $k = 1, 10$ and *free* k (k is cross-validated). For the ablation studies, we fix ground truth boxes as object proposals and report recall on the **Phrase Classification** task, which masks errors from detector. We report our ablation studies on VG150 (Xu et al. 2017) dataset.

Results

Table 1 shows results on VRD dataset. The quality of bounding box proposals from the detector have significant effect on relationship recall results. For a fair comparison, we divide the results in two parts based on the detection box proposals used. We compare previous works which use test set detection proposals from (Lu et al. 2016) and also report results with improved proposals from the detector, generated in a manner similar to (Zhang et al. 2019a).

Our method shows significant improvement over the prior state-of-the art methods with proposals from (Lu et al. 2016). On relationship detection, there is 5% point increase over the previous best results. These results are equally well translated to the task of phrase detection where our model is able to achieve nearly 8% points improvement. As the proposals used are same, we can infer that the improvement is directly from better relationship recognition. With improved proposals, there is further improvement in the score.

The results on Visual Genome is shown in Tables 2 and 3. It is not clear value of k used in prior works for VG200 dataset. Hence we report our results for $k = 1$ and free-k. Our method performed best in both the tasks in both Recall-50/100. With k=1, we outperform the most recent state-of-the-art on VG200 dataset (Zhan et al. 2019). by a margin of 2.3%. and with cross validated k, we achieve 4.5% improvement over state-of-the-art for Relation detection task at Recall-50. Similarly, there is significant improvement in Phrase Classification for VG150 split. It should be noted that in both datasets, Recall-100 has significantly higher score. This indicates that multimodal distribution is perhaps better captured with our model as our model optimizes to push higher scores for multiple valid relations which is reflected in Recall score with $k > 1$. It should be noted that our method performed very well for the main evaluation metric that we are interested in, where multiple predictions from each bounding box pair are allowed, supporting our claim of capturing multimodal distribution of triplets.

Ablation Study and Analysis of Results

The most distinct part of our model is the mixture distribution model. As the representative power of mixture models increases with increasing mixing components, we first evaluate our model with varying Rank of tensor decomposition or number of mixing components of the model. To evaluate the effectiveness of each of these components, we report **phrase classification** results on VRD and VG150 datasets. Phrase classification isolates the factor of object localization accuracy by using ground truth boxes, meaning that it focuses more on the relationship recognition ability of a model. The first 5 models are tested without the dataset prior

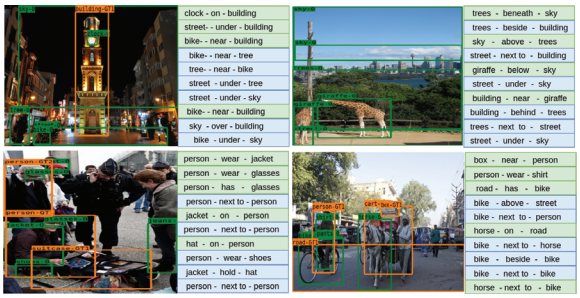

Figure 4: Most probable results predicted by our model. Green shade indicates a match with ground truth labels. Matched results indicate strong presence of multimodality and the triplets generated that do not match are mostly reasonable.

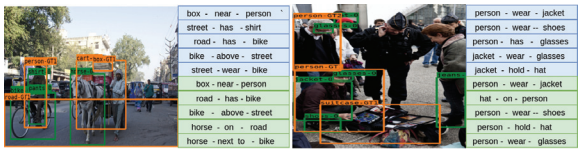

Figure 5: Results of model with prior indicate it helps in removing spurious triplets. without prior with prior

or X_{sel} variable to analyze the relative gains of increasing the number of mixing components. We further evaluate the effect of X_{sel} variable and dataset prior factor with addition to Rank-5 model. The ablation results are shown in Table 4. Both the datasets show significant improvement when the rank is increased from rank-1 to rank-5. Note that rank-1 tensor decomposition is equivalent to independent distributions of triplets and rank-5 to a mixture of 5 components. From rank-1 to rank-5, recall@50 score with multiple predictions per box pair increases by 5% points in the VG dataset and by 2% points in the VRD dataset. This shows that tensor decomposition structure is better able to capture the multimodal triplet distribution. Clearly, gain in score with multiple predictions ($k > 1$) is better than with single prediction ($k = 1$) and with recall@100 is better than recall@50 across datasets. This result is in line with our motivation of optimizing the model for multiple relations instead of one. With a rank-1 independent distribution, the model is optimized for a single top-prediction, hence recall@50 with $k = 1$ has similar score across ranks 1 to 5. With mixture distribution the model is optimized for a set of valid predictions, hence recall@50/100 score for $k > 1$ increases substantially with increasing rank. As we further increase the mixing components, there is a small reduction in score. From this, we can infer that most conditional triplet distributions have a small number of modes and increasing the rank further just increases number of parameters in the model. Further including the X_{sel} variable and dataset prior to the Rank-5 model improves the score substantially supporting our assumption of strong bias towards a small set of relationships.

Dataset	Ablation model	Phrase classification			
		mult preds (free k)		single pred (k=1)	
		R@50	R@100	R@50	R@100
VG150	Rank-1	37.35	44.49	32.16	34.30
	Rank-2	42.53	47.52	33.64	35.26
	Rank-3	41.89	48.35	**33.89**	35.31
	Rank-4	41.74	48.22	33.29	35.14
	Rank-5	**42.71**	**48.95**	32.21	**35.35**
	w/ X_{sel}	46.25	53.05	34.57	36.57
	w/ X_{sel} & prior	**47.54**	**54.69**	**35.60**	**37.68**
VRD	Rank-1	39.01	46.3	31.12	33.04
	Rank-2	**41.29**	**51.58**	**33.43**	**35.16**
	Rank-3	40.28	49.5	32.46	34.96
	Rank-4	39.63	49.64	32.03	33.52
	Rank-5	40.47	50.75	32.59	34.06
	w/ X_{sel}	42.89	52.47	33.51	35.78
	w/ X_{sel} & prior	**44.02**	**53.99**	**34.14**	**36.07**

Table 4: Ablation results on the task of phrase classification.

Qualitative Results:

Figure 4 shows some of the qualitative results of our model. From the overlap between ground-truth and predicted labels, it can be seen that the conditional distribution is at least bimodal if not tri-modal. Also, the triplets that are generated by the models but are not annotated are mostly reasonable.

In Figure 5, we visualize results without prior multiplication. We see multiple cases of erroneous phrases such as 'street-has-shirt', 'street-wear-bike', and 'jacket-wear-glasses' that are unlikely to appear in common usage. Interestingly, such spurious triplets are pushed down from the top of the output lists after multiplication with prior. However, non-spurious triplets that do not appear in the training set may also be pushed down; using external language datasets may improve performance by providing improved usage prior.

Conclusion

We observe that the conditional distribution of relation triplets given input bounding box pair in relation detection tasks is often multimodal. We propose use of mixture of rank-1 tensors for modeling the conditional distribution. This enables the model to capture multimodal properties of the distribution with a reasonably small number of model parameters while being efficiently trainable. We further model the generative labeling process to handle missing annotations and remove spurious triplets with principled incorporation of dataset prior. We show that each of these improve performance on the task of visual relationship recognition. Further improvements may include a language prior from external datasets and with our tensor-decomposition model, it should be possible to do graph inference with higher-order triplet potential over all the boxes in the image.

Acknowledgements

This work is supported by NUS AcRF Tier 1 grant R-252-000-639-114.

References

Anandkumar, A.; Ge, R.; Hsu, D.; Kakade, S. M.; and Telgarsky, M. 2014. Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research* 15(1):2773–2832.

Dai, B.; Zhang, Y.; and Lin, D. 2017. Detecting visual relationships with deep relational networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 3298–3308. IEEE.

Girshick, R.; Donahue, J.; Darrell, T.; and Malik, J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 580–587.

Han, C.; Shen, F.; Liu, L.; Yang, Y.; and Shen, H. T. 2018. Visual spatial attention network for relationship detection. In *2018 ACM Multimedia Conference on Multimedia Conference*, 510–518. ACM.

Jae Hwang, S.; Ravi, S. N.; Tao, Z.; Kim, H. J.; Collins, M. D.; and Singh, V. 2018. Tensorize, factorize and regularize: Robust visual relationship learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1014–1023.

Janzamin, M.; Sedghi, H.; and Anandkumar, A. 2015. Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods. *arXiv preprint arXiv:1506.08473*.

Kolda, T. G., and Bader, B. W. 2009. Tensor decompositions and applications. *SIAM review* 51(3):455–500.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123(1):32–73.

Liang, K.; Guo, Y.; Chang, H.; and Chen, X. 2018. Visual relationship detection with deep structural ranking. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Liang, X.; Lee, L.; and Xing, E. P. 2017. Deep variation-structured reinforcement learning for visual relationship and attribute detection. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 4408–4417. IEEE.

Lin, G.; Shen, C.; Van Den Hengel, A.; and Reid, I. 2016. Efficient piecewise training of deep structured models for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3194–3203.

Lu, C.; Krishna, R.; Bernstein, M.; and Fei-Fei, L. 2016. Visual relationship detection with language priors. In *European Conference on Computer Vision*, 852–869. Springer.

Plummer, B. A.; Mallya, A.; Cervantes, C. M.; Hockenmaier, J.; and Lazebnik, S. 2017. Phrase localization and visual relationship detection with comprehensive image-language cues. In *Proceedings of the IEEE International Conference on Computer Vision*, 1928–1937.

Rabanser, S.; Shchur, O.; and Günnemann, S. 2017. Introduction to tensor decompositions and their applications in machine learning. *arXiv preprint arXiv:1711.10781*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 91–99.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Wrigley, A.; Lee, W. S.; and Ye, N. 2017. Tensor belief propagation. In *International Conference on Machine Learning*, 3771–3779.

Xu, D.; Zhu, Y.; Choy, C. B.; and Fei-Fei, L. 2017. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2.

Yin, G.; Sheng, L.; Liu, B.; Yu, N.; Wang, X.; Shao, J.; and Loy, C. C. 2018. Zoom-net: Mining deep feature interactions for visual relationship recognition. *arXiv preprint arXiv:1807.04979*.

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

Zellers, R.; Yatskar, M.; Thomson, S.; and Choi, Y. 2018. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5831–5840.

Zhan, Y.; Yu, J.; Yu, T.; and Tao, D. 2019. On exploring undetermined relationships for visual relationship detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5128–5137.

Zhang, H.; Kyaw, Z.; Chang, S.-F.; and Chua, T.-S. 2017a. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5532–5540.

Zhang, H.; Kyaw, Z.; Yu, J.; and Chang, S.-F. 2017b. Ppr-fcn: weakly supervised visual relation detection via parallel pairwise r-fcn. In *Proceedings of the IEEE International Conference on Computer Vision*, 4233–4241.

Zhang, J.; Kalantidis, Y.; Rohrbach, M.; Paluri, M.; Elgammal, A.; and Elhoseiny, M. 2019a. Large-scale visual relationship understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 9185–9194.

Zhang, J.; Shih, K. J.; Elgammal, A.; Tao, A.; and Catanzaro, B. 2019b. Graphical contrastive losses for scene graph parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 11535–11543.

Zhu, Y., and Jiang, S. 2018. Deep structured learning for visual relationship detection. In *Thirty-Second AAAI Conference on Artificial Intelligence*.