Rethinking Visual Relationships for High-level Image Understanding

Yuanzhi Liang¹,², Yalong Bai¹, Wei Zhang¹, Xueming Qian², Li Zhu², and Tao Mei¹

¹JD AI Research
²Xi’an Jiaotong University
liangyzh13@stu.xjtu.edu.cn ylbai@outlook.com wzhang.cu@gmail.com {qianxm, zhuli}@mail.xjtu.edu.cn tmei@live.com

Abstract

Relationships, as the bond of isolated entities in images, reflect the interaction between objects and lead to a semantic understanding of scenes. Suffering from visually-irrelevant relationships in current scene graph datasets, the utilization of relationships for semantic tasks is difficult. The datasets widely used in scene graph generation tasks are split from Visual Genome by label frequency, which even can be well solved by statistical counting. To encourage further development in relationships, we propose a novel method to mine more valuable relationships by automatically filtering out visually-irrelevant relationships. Then, we construct a new scene graph dataset named Visually-Relevant Relationships Dataset (VrR-VG) from Visual Genome. We evaluate several existing methods in scene graph generation in our dataset. The results show the performances degrade significantly compared to the previous dataset and the frequency analysis do not work on our dataset anymore. Moreover, we propose a method to learn feature representations of instances, attributes, and visual relationships jointly from images, then we apply the learned features to image captioning and visual question answering respectively. The improvements on the both tasks demonstrate the efficiency of the features with relation information and the richer semantic information provided in our dataset.

1. Introduction

Image understanding based on individual instances, e.g., object detection, classification and segmentation, has witnessed significant advancement in the past decade. Usually, a natural image consists of multiple instances in a scene, and most of them are interacting in certain ways. To go a step further, deeper image understanding requires a holistic view of modelling interactions between individual objects. Visual relationship [14, 4, 25, 30, 32], which encodes the interplay between instances, becomes the indispensable factor for high-level image understanding tasks, such as image retrieval [10], captioning [27], and visual question answering [15]. In existing literature, visual relationships are mostly represented as a scene graph as shown in Figure 1 and the relationship is represented to a triplet form as \((\text{subject}, \text{relationship}, \text{object})\) intuitively.

Similar to most deep learning researches, the image datasets serve as the backbone for visual relation understanding. Up to now, Visual Genome (VG) data set [11] provide the largest relationship combinations, and there are nearly 21 pairwise relationships for each image in VG. It offers abundant data for relationships studying. Since the

![Figure 1. A snapshot of scene graph data and its related scene graph tasks with different settings. In different conditions of detection inputs, the scene graph generation task can be divided into Scene Graph Detection (SGDet), Scene Graph Classification (SGCls), Predicates classification (PredCls), and Predicate detection (PredDet).](image-url)
VG150 is a split set of VG, which is constructed by the top 150 objects categories and 50 relation categories in Visual Genome. 

VG150 is a split set of VG, which is constructed by the top 150 objects categories and 50 relation categories in Visual Genome. 

All these facts reveal that current studies in relationships get stuck and relation information in images has not been fully explored. 

To avoid above defects from relation data and construct a better dataset, we propose a novel method to automatically discriminate visually-relevant relationships and construct a new scene graph dataset named Visually-relevant Relationships in Visual Genome (VrR-VG). Figure 2(b) shows a scene graph snapshot in VrR-VG.
Figure 3. The top frequency relationships in VG150 based on statistical frequency. The top 16 relationships take 93.56% of all the relation labels.

Figure 4. Statistical-biased relationships in VG150. (a) shows the proportions of relationships when subject instance is “man” and object instance is “nose”. (b) shows the proportions when subject is “man” and object is “jacket”

Table 1. Comparison between existing visual relationship datasets.

| Dataset         | # Objects | # Relations | # Images   |
|-----------------|-----------|-------------|------------|
| Visual Phrase   | 8         | 9           | 2,769      |
| Scene Graph     | 266       | 68          | 5,000      |
| VRD             | 100       | 70          | 5,000      |
| Open Image      | 57        | 10          | 1,743,042  |
| Visual Genome   | 33,877    | 40,480      | 108,077    |
| VG150           | 150       | 50          | 87,670     |

VrR-VG is more balance and contains more visually-relevant relationships as shown in Figure 5. To demonstrate the difficulty of our VrR-VG, we report the performance of several methods in scene graph generation [30]. The experiments show significant performance decrease in all the metrics when adopting our data splits. We also prove that barely frequency analysis no longer works in VrR-VG. Moreover, the features trained from VrR-VG have convincible results in some text-image multimodal application like Visual Question Answering (VQA), which also indicates that more complex semantic information contained in our dataset than VG150. All of these results provide further confirmations of the defects in previous scene graph datasets and also exposes insufficiency in current relationships research works.

The main contributions of this paper are summarized as follows:

1. A novel method for selecting visually-relevant and valuable relation label is proposed. Both of the Positional Relationships and Statistically Biased Relationships are detected by using this method. Complex and visually-relevant relationships are reserved. Depend-

2. A feature embedding method trained with location, categories, attributes of single instance, and the relationships among instances. With the proposed method, more semantic information in relationships is added into features.

3. The performance improvements in VQA task and more complicated predicates appeared in caption result demonstrate that VrR-VG contains more valuable visual relationships than previous scene graph dataset, and the visual representations learned on our dataset have a higher ability in semantic expression.

2. Related Work

Datasets We list all relationship related datasets in Table 1. Visual phrase dataset [20] focused on relation phrase recognition and detection. In the visual phrase, a relation triplet is annotated as a whole target. Visual phrase dataset [20] contained 8 objects categories from Pascal VOC2008 [6] and 17 relation phrases with 9 different relationships. Scene graph dataset was proposed in [10] and mainly explored the ability of image retrieval by scene graph. The VRD dataset [14] intended to benchmark the relationship detection task, and contains 37,993 relation triplets with 6,672 unique triplets Open image [26] provided the largest amount of images. It was a brilliant resource for object detection and also presented a challenging task for relation-
ship detection. PIC [1] is the newest among all the datasets we listed. It proposed an interesting task for combination the instance segmentation and relationships. In stead of taking detection bounding boxes and classes as supervision, the segmentation masks of instances became the main goal in PIC. Moreover, UnRel [17] dataset focus on some rare relationships and present a new task for relation presentation in weakly-supervision. While due to the main target of UnRel is representing uncommon relationships, the amount of annotations is limited. There are only 1071 images in UnRel dataset, which are hard to be adapted in a supervised task.

Visual Genome (VG) [11] had the maximum amount of relation data with the most diverse objects and relation categories in all listed datasets. As shown in Table 1 millions of object labels, relation triplets were contained in VG. However, the relations in VG are extracted from image captions, which contains lots of noisy and duplicate relations. Thus VG150 is constructed by processing VG dataset and only high frequency relationships are kept. However, most of high frequency relationships are visual irrelevance as we mentioned before.

Representation Learning Numerous deep learning methods have been proposed for image representation learning. These methods offer two aspects in image understanding, which representation with single instance and multiple instances. In single instance, GoogLNet [22], Resnet [8], Inception [21], ResNext [24], etc. are trained with Imagenet [5] dataset and focus on the single instance classification. Since the supervise labels are instance categories, the methods tend to give a holistic representation of images and figure out the features with the single salient instance attentioned. Furthermore, as more than one instances exist in the images, focusing on one salient instance is not enough to represent the scenes in images. To explore multiple instances, methods in detection task provide some inspirational thoughts. Jin et al. [9] adapt selective search [23] to give salience region proposals. The similar idea also appears in RCNN [2], in which the network achieve many region proposals first and work out detection result for every instance. Faster-RCNN [13] further improved the idea of region proposals and provide a faster and more elegant method to limited region proposals. Established by Faster-RCNN’s region proposals, Peter et al. [2] proposed a bottom-up and top-down attention method to represent images. They utilize the locations, categories and attributes of instances to learn the representation and get improvement in some semantic tasks. In our work, we go deeper in multiple instances representation by adding inter-instance information in features. All the isolated instance information like locations, categories, attributes, together with instances relationships are all fully utilized in representation learning.

3. Visually-relevant Relationships Dataset

In this section, we introduce a novel visually-relevant relationship discriminator (VR²D). VR²D is a simple fully connected network aims to recognize relation labels directly by entities classes and bounding boxes. Since there no visual information feeding to VR²D, relationships have high predication accuracy would be tend to regard as visually-irrelevant relationships. After filtering out these visually-irrelevant relationships and reducing duplicate relationships by hierarchical clustering, we constructed a new dataset named Visually-relevant Relationships Dataset (VrR-VG) from original VG dataset.

3.1. Visually-relevant Relationship Discriminator

To distinguish visually-irrelevant relationships, a hypothesis is proposed that, if a relationship is predictable according by any information except visual information, the relationship is visually-irrelevant. In our work, a simple visually-relevant relationship discriminator (VR²D) is proposed for selecting relationships. To prevent the influence of overfitting caused by large amount of parameters, the network structure design follows the guideline “tinier is better”. Our VR²D aims to recognize relationships without providing any kinds of visual information. Specifically, the inputs are word vectors of instances’ category learned from Natural Language corpus and the coordinates of the corresponding bounding boxes. Glove [16] is adapted for word
vector.

Each bounding box of instance in the image can be defined by a four-tuple \((x, y, h, w)\) that specifies its top-left corner \((x, y)\) and its height and width \((h, w)\). Here we denote the position embedding for subject and object as \(p_o\) and \(p_s\) respectively, where \(p_o = \{x_o, y_o, h_o, w_o\}\) and \(p_s = \{x_s, y_s, h_s, w_s\}\). The bounding boxes set of given object and subject in related entities are embedded to a jointly vector as following equation:

\[
p_j = \left[ a_x, a_y, w_o, w_s, h_o, h_s, c_s - c_o, c_o - w_s, h_s - h_o, \right.
\]
\[
\left. \frac{(c_s - c_o)}{w_s}, \frac{(c_o - c_s)}{h_s}, \log(\frac{w_o}{w_s}), \log(\frac{h_o}{h_s}) \right]
\]

where \(a_x, a_y\) are offsets of boxes computed by the difference between the coordinates of subject and object, \([w_o, h_o]\) and \([w_s, h_s]\) are width and height of bounding box of subject and object respectively, and \([c_x^o, c_y^o]\) and \([c_x^s, c_y^s]\) are the coordinates of the center of the corresponding boxes. Above position embedding values provide position information for our network.

The network details are given in Figure 6 where \(v_o\) and \(v_s\) are the word vectors of subject and object categories. \(W^D\) are learnable weights in VR\(^2\)D network. The word vectors of object and subject are processed by a simple fully-connected layer. Then, the output features are concatenated with position embedding \(p_o, p_s\) and \(p_j\). Finally, another two fully-connected layers and batch normalization layers are applied for classifying relation labels. We discard relationships which have larger accuracy than a threshold \(\alpha\), and those reserved relationships are selected for generating the dataset. In this paper, we set \(\alpha\) as 50% due to the trade-off between dataset scale and visually-relevant semantic quality.

The VR\(^2\)D merely contains three fully-connected layers, but it is already sufficient to predict most of the visually-irrelevant relationships like “wear”, “on”, “above”, etc. It is worth noting that more than 54% of relation labels in VG150 can be predicted with at least 50% accuracy by using such a rude neural network without any visual information.

### 3.2. Dataset Construction

We pre-process VG and extract top 1600 objects and 500 relationships to generate a basic data split. The raw relation labels in scene graph contain many duplications, such as “wears” and “is wearing a”, “next” and “next to”. Those labels may confuse network because all those labels are correct to the same object and subject combination. We represent the labels by Glove word vector, and then cluster the vectors by the hierarchical method. This simple operation reduces label categories from 500 to 180, which is more diverse than the formal 20 relations in VG150 and reserve sufficient semantic level information. Then, to exclude visually-irrelevant relationships, the VR\(^2\)D network is utilized to train and evaluate with the 180 clustered labels. Finally, we get 117 relation labels as VrR-VG relationships.

In the comparison of image amount, the VG150 data split, which was adapted in many relationships representation works [30, 25], contains 87670 images and 588586 triplet pairs. VrR-VG has 58983 images and 23375 relation pairs, which is less than VG150. However, as shown in Figure 7, the distribution of our dataset is more balance. Since VG150 is generated depending on the label frequency, the easy and general relationships are inevitably involved in the dataset. Considering the data bias in original Visual Genome, the imbalance data distribution inherits when generating data split for both of VG150 and VrR-VG.

Moreover, labels like “on”, “of”, etc. in VG150 are too trivial to describe entities interaction, and there are also some relationships having same semantic meaning appear in the top relationships in VG150, e.g. “wears” and “wearing” are regard as two different relationships and accounting for 11.87% and 0.84% in VG150 respectively. Comparatively, our top12 labels are more significant in the semantic domain. Relationships like “hanging on”, “playing with”, etc. are hard to be estimated without enough understanding in corresponding scenes. So, our data split is much more difficult in semantic representation in visual. We also show some exemplar scene graphs with randomly sampled images in Figure A.1. It can be found that, compares with previous VG150 dataset, the relationships in our dataset are more diverse and contain more semantic information for describing the scene.

### 4. Informative Visual Representation Learning

As shown in Figure 8 to modeling entire visual information in an image, the properties of isolated objects like category, position, attribute and the interactions among ob-
5. Experiments

In this section, we discuss the properties of our data split in two sides. One is dataset comparison on traditional scene graph generation task. The other is dataset quality evaluation by applying the visual representations learned from different datasets on semantic or reasoning task like visual question answering and image captioning.

5.1. Scene Graph Generation

To estimate the relationships reserved in our split, we compare VrR-VG dataset with other relation datasets on scene graph generation task. As we known, scene graph generation needs to learn semantic in images and infer the connections of entities in semantic level. We mainly evaluate and discuss the latest or widely used scene graph generation methods, including MSDN [12], Vtrans [31], Neural-Motifs [30] and Message-Passing [25] in our experiments. Additionally, Neural-motifs, Message-passing use pre-trained Faster RCNN with VGG backbone as object de-
Table 2. Comparison of methods in relation representation. R50 and R100 are the metrics for evaluating relation detection and scene graph generation. R-gap indicates the difference between R100 and R50. The MSDN and Vtrans methods are evaluated in the other data splits, which are also split from Visual Genome by frequency. While neural-motifs and message passing methods use the same VG150 data split. Additionally, evaluate details about SGCls and PredCls in MSDN and SGCls in Vtrans are not released. As the author acknowledged, the best results of MSDN in SGCls and PredCls are not reproducible. So some numbers are not reported in our experiments.

With the experiments in some relation representation methods, as shown in Table 2, 3, and 4, the performances apparently decrease when adapting our data split. With the relationships selected by our method, the relation representation task becomes more difficult and challenging.

We also evaluate the performance of frequency-based method [30] which uses pre-trained detector in neural-motifs method and then predicts relation labels by statistic label probabilities in training set without visual images inputs. As shown in Table 3, our dataset efficiently reduce the impact of language prior and data bias. The frequency-based method does not practicable any more. In the previous relation dataset VG150, without images input, the statistical method can get similar results to the methods without visual information. Results only decline 3.7, 3.4, and 5.3 in R50 of scene graph detection (SGDet), scene graph classification (SGCls), and predicate classification (PredCls). However, in our data split, 14.7, 23.9, and 23.4 decents appear in the performance of the statistical method, which is far from the models learned from visual information. It indicates that the margin between visual based and non-visual based method have much bigger performance gap on our VrR-VG dataset than is previous widely used VG150 dataset. The performance of the frequency based method is efficiently inhibited in our dataset. The true abilities of semantic understanding can be shown with our dataset.

Further, the results of the frequency based method also reveal that the location information can also be prior to relationships. In VG150, with bounding box ground truth as inputs, the SGCls has 8.9 and 6.4 improvements in R50 and 100, which indicates that the entities positions also contain bias and can help models estimate relation labels without visual information. While, in our data split, performances of frequency based method decease after appending location of entities, which means our data split eliminates positional influence and enforce the models insensitive to object detection.

Moreover, the predicate detection task (PredDet), as the easiest task in relation representation with related paired entities and detection ground truth, stands for the maximum performance. The MAP of the detector is 20.0 and 8.2 in VG150 and our VrR-VG at 50% IOU.

Table 3. Comparison of Frequency based Methods.

Table 4. Comparison of performance in predicate detection.
### Table 5. Comparison between different methods based on image features learned from different datasets for open-ended VQA on the validation split of VQA-2.0 dataset.

| Method | Used Relation | Dataset | Yes/No | Numb. | Others | All |
|--------|---------------|---------|--------|-------|--------|-----|
| Bottom-Up [2] | X | Bottom-Up | 80.3 | 42.8 | 55.8 | 63.2 |
| Resnet [2] | X | Bottom-Up | 77.6 | 37.7 | 51.5 | 59.4 |
| MUTAN [3] | ✓ | Bottom-Up | 81.90 | 42.25 | 54.41 | 62.84 |
| | X | VG150 | 79.00 | 39.78 | 49.87 | 59.49 |
| | ✓ | VrR-VGobj | 80.46 | 42.93 | 54.89 | 62.93 |
| | ✓ | VrR-VG | 83.09 | 44.83 | 55.71 | 64.57 |
| MFH [29] | X | Bottom-Up | 82.47 | 45.07 | 56.77 | 64.89 |
| | ✓ | VG150 | 78.86 | 38.32 | 50.98 | 59.80 |
| | X | VrR-VGobj | 82.37 | 45.17 | 56.40 | 64.68 |
| | ✓ | VrR-VG | 82.95 | 45.90 | 57.34 | 65.46 |

#### 5.2. Representations Learning for Semantic Tasks

To evaluate the relation quality in semantic level, we choose VQA and caption tasks in experiments and adapt our split to inspect semantic improvements of the dataset. We also compared our informative visual representation learning method with traditional visual representation learning method, the experimental results demonstrate that the visually-relevant relationship play a important role in high-level visual understanding tasks. For the implementation of our proposed informative visual representation method, following the settings of previous bottom-up and top-down method [2], we set $k = 32$ for all experiments in this section and image features are integrated into 2048 dimensions by mean-pooling. Additionally, we introduce a new dataset VrR-VGobj, which is generated by excluding relation data from VrR-VG, for ablation study. We apply our proposed feature learning for VrR-VGobj too, but without the weight of the relationship predication loss is set as 0.

**VQA** We applied two widely used VQA methods MUTAN [3] and MFH [29] for evaluating the quality of image feature learned from different datasets. Table reports the experimental results on validation set of VQA-2.0 dataset. We can find that features trained with our VrR-VG obtain the best performance in all the datasets. We also compared the dataset used in Bottom-Up attention [2], which is regard as the strongest feature representation learning method for VQA.

In details, with relation data, our complete VrR-VG performs better than dataset used in Bottom-Up attention and VrR-VGobj. The results indicate that the relation data is useful in VQA task, especially in the highly-semantic related questions. It also demonstrate that our proposed informative visual representation method can extract more useful relationships.
features from images. Besides, we also apply our proposed feature learning method on VG150 dataset. Since VG150 contains too many visually-irrelevant relationships which can be inferred easily by data bias as we mentioned. As a result, the features learned from VG150 usually lack of the ability to represent complex visual semantics. The higher quality of relation data energize the features learned from our dataset and lead to a better performance in the open-ended VQA task.

**Image Captioning** In this part, we adapt our constructed dataset VrR-VG and our proposed entity, attributes, localizations and interactions jointly feature representation learning method to image captioning task. Similar with the experiment process used in VQA task, we first generate the image features based on VG150, VrR-VG_{obj}, and VrR-VG respectively. Then we apply the caption model proposed in [2] for these image features with same settings in fair.

The experiments results are shown in Table 6. As shown in the table, we report the performances in our split and VG150 in both original optimizer for cross entropy loss and CIDEr optimizer for CIDEr score. Features generated from our data split works better then VG150. All the metrics in caption have better performance when using both of two different optimizers. Moreover, in the comparison of adding relation or not, Our complete VrR-VG has better performance than the VrR-VG_{obj} which only contains objects information in original VrR-VG. This indicates that the relation information is utilized and useful. The improvement shows the efficiency of our dataset and also reveal the power of relationships in semantic tasks.

Specifically, to show the superiority of the relation data in our split, as shown in Figure 10 some examples are given to show the difference of features in VG150 and VrR-VG. To fairly compare, both caption examples have the accurate recognition in objects. We focus on the difference of predicates, which can better reflect the influence of relation data. In examples of caption results, our features tend to have more diverse predicates and more vivid description. Rather than some simple predicates like “on”, “standing”, etc., our features provide more semantic information and help models achieve more complex expression like “writing”, “grazing”, etc.

### 6. Conclusion

In this paper, we design a tiny network as relation discriminator to select visually-relevant relationships. With the selected relationships, a visually-relevant relationships split in Visual genome is generated in our work. Compared with original VG splits, this split contains more valuable relationships and more semantic information, which are hard to estimate merely by statistical bias or detection ground truth. We also proposed an informative visual representation learning method which is designed to learning image feature according to entities’ label, localization, attributes and interactions among entities jointly. The significant improvements on VQA and image captioning task demonstrate that: (1) our constructed dataset VrR-VG has better quality than previous datasets, (2) visual relationship information is helpful for high-level semantic understanding tasks, (3) our proposed informative visual representation learning method can effectively model different kinds of visual information jointly.

### References

[1] picdataset.com. [http://picdataset.com/challenge/index/](http://picdataset.com/challenge/index/), 2018.

[2] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pages 6077–6086, 2018.

[3] H. Ben-younes, R. Cadène, M. Cord, and N. Thome. MUTAN: multimodal tucker fusion for visual question answering. CoRR, abs/1705.06676, 2017.

[4] B. Dai, Y. Zhang, and D. Lin. Detecting visual relationships with deep relational networks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017.

---

| Methods       | Used Relation | BLEU-1 | BLEU-4 | METEOR | ROUGLE-L | CIDEr | SPICE |
|---------------|---------------|--------|--------|--------|----------|-------|-------|
| SCST [19]     | ×             | -      | 30.0   | 25.9   | 53.4     | 99.4  | -     |
| LSTM-A [28]   | ×             | 75.4   | 33.2   | 26.9   | 55.8     | 108.8 | 20.0  |
| Cross-Entropy Loss | Resnet       | ×      | 74.5   | 33.4   | 26.1     | 54.4  | 105.4 | 19.2  |
|               | VG150         | ✓      | 74.2   | 32.7   | 25.3     | 53.9  | 102.1 | 18.5  |
|               | VrR-VG_{obj}  | ×      | 76.2   | 35.4   | 26.8     | 55.7  | 110.3 | 19.9  |
|               | VrR-VG        | ✓      | 76.9   | 36.0   | 27.2     | 56.3  | 114.0 | 20.4  |
| CIDEr Optimization | Resnet       | ×      | 76.6   | 34.0   | 26.5     | 54.9  | 111.1 | 20.2  |
|               | VG150         | ✓      | 76.7   | 32.7   | 25.8     | 54.3  | 108.0 | 19.6  |
|               | VrR-VG_{obj}  | ×      | 78.8   | 35.8   | 27.3     | 56.4  | 116.8 | 21.0  |
|               | VrR-VG        | ✓      | 79.4   | 36.5   | 27.7     | 56.9  | 120.7 | 21.6  |

Table 6. Experimental results in image captioning task.

---

4329
4330
A. Appendix

A.1. Scene Graph Data Comparison

Some scene graph examples are given here. In VrR-VG, only visually-relevant relationships are contained. As shown in figures, scene graph generated from VrR-VG tend to have more complicate semantic expressions which can offer more diverse and sufficient information for multiple tasks.

A.2. Image Captioning Results Comparison

In this section, as shown in Fig[10], we provide more caption results to show the difference between features learned from VG150 and ours. With more visually-relevant relationships, the predicates in captions become more diverse.
Figure 10. Examples of caption results.