Focusing Visual Relation Detection on Relevant Relations with Prior Potentials

François Plesse, Alexandru Ginsca, Bertrand Delezoide
DRT LIST DIASI LASTI
{francois.plesse,alexandru.ginsca,bertrand.delezoide}@cea.fr

Françoise Prêteux
Ecole des Ponts - Cermics
francoise.preteux@enpc.fr

Figure 1: Image examples from Visual Genome [12]. Images with similar sets of objects can have very different meanings, which is why understanding the relations between objects is vital to image understanding.

Abstract

Understanding images relies on the understanding of how visible objects are linked to each other. Current approaches of Visual Relation Detection (VRD) are hindered by the high frequency of some relations: when an important focus is put on them, more meaningful ones are overlooked. We address this challenge by learning the relative relevance of relations, and integrating this term into a novel scene graph extraction scheme. We show that this allows our model to predict relations on fewer and more relevant object pairs. It outperforms MOTIFNET*, a state of the art model, on the Visual Genome dataset. It increases the Class Macro recall, the metric we propose to use, from 38.1% to 44.4%. In addition, we propose a new split of Visual Genome, with a more balanced relation distribution, emphasizing on the detection of uncommon relations and validates the use of the previous metric. On this set, our model outperforms* MOTIFNET *on all metrics,* e.g. *from 39.6% to 44.0% at 10 predictions per image on the relation classification task.*

1. Introduction

Images are more than the sum of their parts and only knowing what objects are visible is not sufficient to fully understand their content. Indeed, the relations between objects shape the understanding of an image, because they make an image stand out and bring meaning that is missing when considering only objects. Consider images in Figure 1. Knowing that the images pictures a horse, rider and water does not capture the meaning of the image. However, the relation *'rider falls from horse'* shows that the first image depicts an unusual event, because usually, riders tend to stay on horses.

The broad problem of extracting meaning from images has been tackled in recent years. First, the foundations of image recognition were laid with successful neural architectures [23, 6] and soon after with new architectures for object detection [3, 22, 21]. This enabled the extraction of higher order concepts, such as image caption generation [11, 25, 28], which lays at the intersection between image analysis and natural language processing. In this work, we focus on the detection of visual relations, aiming to represent an image by its scene graph, i.e. representing visible objects by nodes and relations between them by edges of the graph. These relations can be interactions or spatial relations (e.g. *fall from* between *rider* and *horse* in Figure 1).

The task of extracting a scene graph is a combinatorial problem, because the number of object pairs increases quadratically with respect to the number of objects and each pair can usually be described by several relations. Furthermore, images are very rich sources of information, thus it is necessary to filter out information in order to keep a low volume of stored data, low bandwidth use and noise levels; and most of the available information is not relevant and can detract from relations with higher information content. Moreover, as pointed in [16], much of this information is not relevant to humans, because it is redundant with their prior knowledge. Thus people tend not to mention these pieces of information when describing images, omitting attributes that are "obvious or typical". For example, in Figure 2, several hundred relations are true, among which *bed in front of wall, bed near wall, doctor has jacket, doctor in shirt, picture below picture, pen hanging from jacket*. However, these relations are not mentioned in annotations of Visual Genome [12].

For this reason, it is necessary to model the relevance of relations: this characterizes how salient they are and

Figure 2: Image from Visual Genome. Several hundred relations are true but provide little understanding of the image.

how much information content they represent. Modelling the relevance of relations increases, without loss of performance, the variety and relevance of predicted relations. This change to the generation of scene graphs contrasts from most recent works which focus on the training process of relation classifiers; however it can have a strong impact on the model performance, especially for small scene graphs.

Finally, many recent models mainly focus on a high retrieval rate of relations per image. However, the most studied VRD dataset, Visual Genome, has a highly skewed relation distribution towards a few relations, such as 'on' and 'wear' (in respectively 29% and 12% of the training relations). Thus, the overall recall provides a limited picture on the capabilities of the model. For tasks focusing on a single image, increasing the overall recall is important. However, when considering a high number of images, the rarer relations are critical to differentiate one image from another. Moreover, many applications require learning new classes, with little available data. Our method is especially useful for those difficult cases. To better evaluate the capacity of models to predict unusual and relevant relations, we propose a new split of Visual Genome (VG) [12] and an additional metric aimed at showing how the model performs at retrieving rarer relations.

Contributions Our contributions are summarized as follows:

1. *A Relation Detection Model with Relevance*. To filter out irrelevant and obvious relations, we predict relation relevance at test time, making use of dataset statistics to decrease noise and increase precision.

2. *A new evaluation metric* Due to the aforementioned skewness of relations, the overall recall on the test set only gives limited information on the capability of the model to generalize to new situations, especially less common relations. To remedy this, we propose a new evaluation metrics which increases the importance of retrieving uncommon relations.

3. *A new split of Visual Genome* The most studied split of VG is highly skewed towards a small number of relations, due to the annotation process and the preponderance of similar scenes such as streets. This is a hindrance to the evaluation of VRD models as predicting the most common relation for two objects is a viable strategy in many cases. To offset this, we propose a split where the number of relation examples for each object pair is more evenly distributed.

2. Related Work

Visual Relation Detection has been shown to be beneficial to image generation [9], image retrieval [10] and Visual Question Answering [24]. Hence, it has recently received increased attention, first by focusing on $(human, action, object)$ triplets in images [2, 4], but also on the broader task of detection and classification of (*subject, relation, object*) triplets. These models are based on 3 components: (1) an object detector [22], (2) a model that extracts spatial features from coordinates and (3) a three-branch classification model: one for the classification of each object of the relational phrase and one to classify relations. Many recent works focus on taking advantage of statistical dependencies between relational phrase constituents [13, 27, 30, 17, 14, 29, 32, 33], leveraging semantic knowledge [15, 30, 14, 19, 29] and spatial information [20, 18, 30, 35, 32, 33].

Zellers et al. [32] use bi-directional LSTM networks [8] on the visual features of the detected objects and object pairs in order to aggregate the global image context and make use of the dependencies between all constituents of the image. Woo et al. [26] integrate contextual information using a relational embedding. Object features are refined using an attention mechanism over all objects, called a relational embedding. They show that this embedding coincides with ground truth relations and represents the interdependencies between objects. Our work contrasts with these because we use an estimated relation relevance to extract graphs focused on important relations.

Concurrently, datasets providing object bounding box annotations and relations for pairs of objects have been released, allowing the training of visual relations detection models [15, 12, 2] and to test them on seen, unseen [15] or unusual relations [18].

Concept Relevance refers to the phenomenon whereby the probability that a concept is visible differs from the probability that it is annotated by a human. Berg et al. [1] showed that objects that are small or far from the center of an image are less likely to be mentioned. Unusual objects and people however, tend to be mentioned more often. Misra et al. [16] tackle this discrepancy by separately modelling the presence of an object and its relevance so that the model may simultaneously predict a high probability of

presence and a low relevance. Here, we learn the relevance of relation from its representation.

Model Bias Highly skewed datasets, comprised of a high number of a small set of classes, can be a hurdle, preventing the model from correctly classifying instances into the less frequent classes. This can also prevent the evaluation of existing models, which learn to exploit biases and not be penalized for it. Goyal et al. [5] propose an augmented Visual Question Answering test set in which answers to different types of questions have several different answers. They show that most existing works perform much worse on this de-biased dataset. Similarly, to prevent models from exploiting biases in questions, Zellers et al. [31] propose a new VQA dataset in which a justification is asked for the selected answer. We take a similar approach to [5] without additional images, selecting test images to increase the diversity of relations for each pair of object categories.

3. Problem definition and notations

Let $\mathcal{D} = \{I_1, \ldots, I_{n_\mathcal{D}}$ a set of images each annotated with n_I object bounding boxes: $\{b_1, \ldots, b_{n_I}\}$.

Let

$$\mathcal{C} = \{o_1, \ldots, o_{n_\mathcal{C}}\}\text{: a set of object classes.} \quad (1)$$

$$\mathcal{R} = \{r_1, \ldots, r_{n_\mathcal{R}}\}\text{: a set of relation classes.} \quad (2)$$

Visual Relation Detection (VRD) is a task whereby a scene graph $G = (V, E)$ is extracted from an image I with

- a set of nodes $V = \{v_1, \ldots, v_{n_I}\}$ where for each i, $v_i \in \mathcal{C}$
- a set of edges $E = \{e_{h\to t} | h \neq t \in [1 \ldots n_I]\}$ where for each h, t, $e_{h\to t} \in \mathcal{R}$

For image I and $h, t \in [1 \ldots n_\mathcal{I}]$, we define V_h, V_t, $R_{h\to t}$ and $Z_{h\to t}$ random variables with values in $\mathcal{C}$, $\mathcal{C}$, $\mathcal{R}$ and $[0, 1]$.

4. Focused VRD with Prior Potentials

We present our model for Visual Relation Detection: FOCUSEDVRD. Since in many images, many true relations are very typical or link unimportant objects, most of them should not be mentioned. Following this observation, we introduce the relevance variable, which allows the model to focus on a smaller number of object pairs and extract more relevant relations from an image. Thus, we propose the following decomposition of scene graph probability:

$$P(G) = \prod_h P(V_h) \prod_{t\neq h} P(R_{h\to t}, Z_{h\to t} | V_h, V_t) \quad (3)$$

This formulation takes into account the variable corresponding to the presence of a relation $R_{h\to t}$ and its relevance to a human observer $Z_{h\to t}$, which we assume are independent variables. In this Section, we describe how the relation distribution $P(R_{h\to t}|V_h, V_t)$ and relevance distribution $P(Z_{h\to t}|V_h, V_t)$ are modelled.

4.1. Network Architecture

First, we extract the visual and spatial representations of relations using a Convolutional Neural Network (CNN). This network first extracts the representation of the image and its corresponding object region proposals as displayed in Figure 3. (a) For each pair of objects, region representations are extracted from the image feature map using ROI-Pooling [22]. (b) Following [32], we add the visual representation of the union bounding box to spatial features extracted from the binary masks using a two-layer CNN. Then the representations of head, tail and union bounding box are passed through two feed-forward layers, noted $\boldsymbol{f}_h$, $\boldsymbol{f}_t$ and $\boldsymbol{f}_{h\to t}$.

4.2. Relation Classification

For each object pair (h, t) (head and tail), the relation probability distribution is computed using

- object feature vectors f_h, f_t
- the visual and spatial feature vectors of the relation: $f^{vis}_{h\to t}$ and $f^{spat}_{h\to t}$

They are passed through feed-forward layers, as shown in Figure 3 (c):

$$\boldsymbol{y}^{vis}_{h\to t} = \boldsymbol{W}^{vis}_h \boldsymbol{f}_h + \boldsymbol{W}^{vis}_t \boldsymbol{f}_t + \boldsymbol{W}^{vis}_e [\boldsymbol{f}_h, \boldsymbol{f}^{vis}_{h\to t}, \boldsymbol{f}_t]$$

$$\boldsymbol{y}^{spat}_{h\to t} = \boldsymbol{W}^{spat}_r \boldsymbol{f}^{spat}_{h\to t}$$

$$p(R_{h\to t}) = \mathrm{softmax}\big(\boldsymbol{y}^{vis}_{h\to t} + \boldsymbol{y}^{spat}_{h\to t} + \log \psi(v_h, v_t)\big) \quad (4)$$

where the weights $\boldsymbol{W}^{vis}$ and $\boldsymbol{W}^{spat}$ project the representations of both objects ($\boldsymbol{f}_h$ and $\boldsymbol{f}_t$) and relation in the $n_\mathcal{R}$-dimensional relation space. All vectors are then summed and the softmax function is applied to the sum to compute the probability over the set of relations, including the *null* relation $\varnothing$. Similarly to the semantic bias used in [32], $\psi(v_h, v_t)$ is a frequency prior computed by measuring the frequency of each relation for each pair of object classes in the training dataset.

4.3. Relevance Estimation with Prior Potentials

In existing datasets, a high number of relations are true in each image, and only a small fraction of them are annotated. Thus supervised models have difficulty extracting relation representations and boundaries that separate relevant and irrelevant relations. Hence, as shown in Figure 3 (d), we propose to use frequency priors in order to smooth the model predictions:

$$P_{\text{human}}(Z_{h\to t}|v_h, v_t) = p_\theta(Z|v_h, v_t) + \phi(v_h, v_t) \quad (5)$$

where $p_\theta(Z_{h\to t}|v_h, v_t)$ is a trained **relevance classifier** and ϕ is a **binary prior potential**.

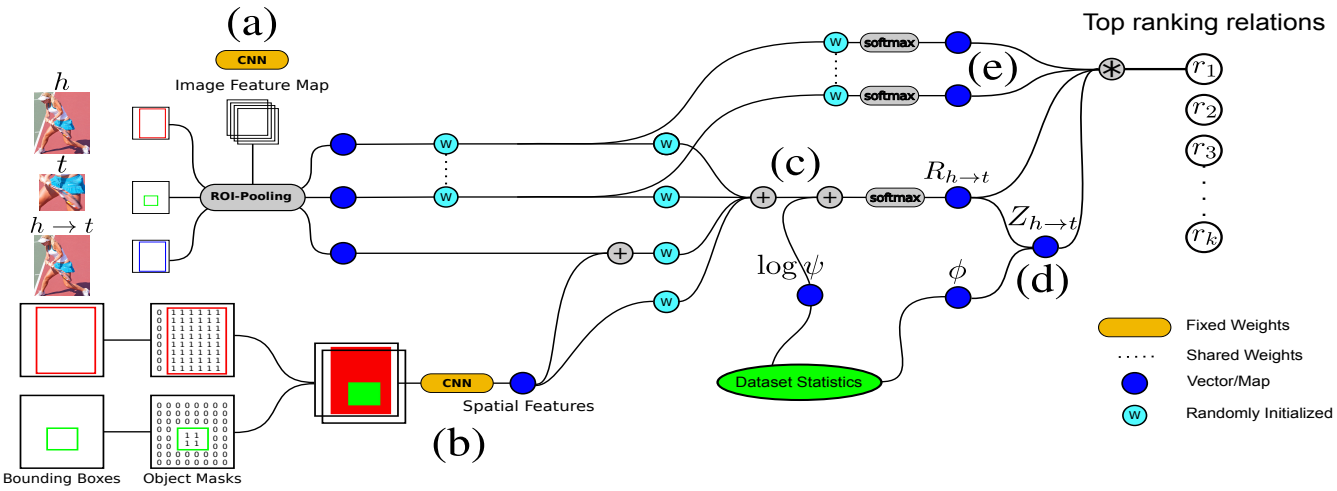


Figure 3: FOCUSEDVRD framework. Visual and spatial representations of each object pair is extracted using CNNs. Statistics-based priors are computed on the training set to predict relations more compatible with common sense (ψ) and improve relevance estimation (ϕ). At test time, a Scene Graph is generated by combining object classification probabilities (e), relation classification probabilities (c) and relevance probabilities (d). This allows the model to focus on relevant object pairs in the image and predict uncommon relations.

Relevance Classifier We model the relevance classification as the probability that any relation is annotated on the given object pair:

$$p_\theta(Z|v_h, v_t) = 1 - p_\theta(R_{h\to t} = \varnothing) \tag{6}$$

Prior Potential of Relevance The relation potentials ψ and ϕ are inspired by [32, 33]. They use a semantic module defined as the empirical distribution of relations given two objects, stating that the number of probable interactions between two object is limited. Similarly, the likelihood that two objects are linked by a relevant relation is estimated by the frequency at which they interact in the training set.

This potential, noted ϕ, is used in addition to the relevance classifier because the relevance of relations in the training set has a high variance which makes the training of this classifier unstable. It is computed by counting the number of co-occurrences of both objects in the train set and the number of times they are in a share relation in the same set.

$$\phi(v_h, v_t) = 1 - P_{train}(R_{i\to j} = \varnothing) \tag{7}$$

$$= \frac{\sum_I \sum_{r\in\mathcal{R}} \mathbb{1}_I(v_h, r, v_t)}{\sum_I \mathbb{1}_I(v_h, v_t)} \tag{8}$$

5. Making the R in VRD matter

Goyal *et al*. [5] observed that the output of Visual Question Answering models was much more dependent on text priors than visual cues on the most commonly studied VQA dataset. In the same vein, we note that, for existing models, relation predictions are conditioned more by the object categories than by their visual relations in a specific image. This is not directly apparent in results as text distributions in Visual Genome [12] are heavily skewed. Figures 4a and 4b display the distribution of relations on Visual Genome for the most frequent pairs of object categories. They are computed by grouping object classes by manually defined hypernyms. They show that the majority relation for each pair often represents from 50% to 75% of the examples. This makes the predictions of models biased towards these relations and the evaluation does not reflect that.

In this section, we propose two ways to highlight the performance of VRD models on rare relations. This is motivated by two observations.

First, many applications require learning new classes, with usually little available data, making these classes rare. Our method is especially useful for those difficult cases. If frequent relations are also important, our method could be combined to another, more adapted for frequent relations.

Second, for many common relations triplets, e.g. (*person, wearing, clothes*) (11% of the training set), we mostly need to detect which objects are related, not to classify the relation. For this task, overlap between objects is enough 90% of the time, according to [32]. Rarer triplets such as (*clothes, hanging from, vehicle*) are important to differentiate images but have a low impact on micro recall. We propose a new split of Visual Genome [12] called VG-RMATTERS, constructed so that predicting the most frequent relation is a less viable solution.

5.1. Dataset Construction

VG-RMATTERS is defined such that predicting the most common relation of each pair of object categories is not a viable strategy. Thus the dataset should be such that the majority relation of each pair of object categories is different between the train dataset and the test dataset. The main difficulty towards this stems from the fact that splits are defined by grouping sets of images together instead of processing arbitrary sets of relations. Thus if an image contains one uncommon relation and a set of common relations, it will offset the distribution of all object pairs in the image.

To this end, we use an algorithm that relies on target numbers of relations for both train and test sets and image scores that quantify how each image impacts those targets.

First, for the test set, the target distribution is defined as

$$p_{test}(e|h,t) \propto \begin{cases} 0 \text{ if } e \text{ is the most frequent class of (h,t)} \\ p_{data}(e|h,t) \text{ otherwise} \end{cases} \tag{9}$$

where (h, e, t) is a relation triplet and p_{data} is the relation probability distribution computed on the whole dataset.

Target counts are then computed as

$$\text{Count}_{test}(h,r,t) = p_{test}(r|h,t) * \text{Count}_{test}(h,t) \tag{10}$$

$$\text{Count}_{train} = \text{Count}_{data} - \text{Count}_{test} \tag{11}$$

where $\text{Count}_{test}(h,t) = 0.1 * \text{Count}_{data}(h,t)$. The proportion of test examples is set to 0.1, against 0.2 for VG-IMP, in order to increase the relation diversity while keeping a comparable number of examples for the rarer relations in the training set. Since we only constrain distributions of object categories separately, the overall relation distribution is similar between both training sets.

From target counts, a score s_I is computed for each image, with higher (resp. lower) scores corresponding to images that should belong to the train (resp. test) set. The split is defined in Algorithm 1. For relation (h, e, t), the relation score is multiplied by the number of instances of this relation, so that relations that impact relation distributions more have a higher absolute value.

5.2. Dataset Characteristics

Table 1 shows the average proportions of majority relation classes and entropies for each pair of object categories for VG-IMP and VG-RMATTERS splits. They show that the diversity of relations for VG-RMATTERS is higher.

Figure 4 shows a comparison of the distributions of relations for the 50 most frequent pairs of object categories in both VG-IMP and VG-RMATTERS. Relations are much more varied in VG-RMATTERS. Differences in distributions shown in Figure 4 and Table 1 suggest that this split better tests how the model is able to recognize relations between two objects by making the most frequent relation a worse option than in VG-IMP.

input : Dataset $\mathcal{D}$ of annotated images
output: Datasets $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$
for *I in* $\mathcal{D}$ **do**
  $s_I \leftarrow 0$;
  for $rel = (h, e, t)$ *annotated in I* **do**
    $\rho_{train}(rel) \leftarrow \frac{\text{Count}_{train}(rel)}{\text{Count}_{train}(h,t)}$;
    $\rho_{test}(rel) \leftarrow \frac{\text{Count}_{test}(rel)}{\text{Count}_{test}(h,t)}$
    $s_{rel} \leftarrow \texttt{Count}(rel) * (\rho_{train}(rel) - \rho_{test}(rel))$
    $s_I \leftarrow s_I + s_{rel}$
  end
end
$\mathcal{D}_{train} \leftarrow 0.9 * |\mathcal{D}|$ images with top s_I scores
$\mathcal{D}_{test} \leftarrow 0.1 * |\mathcal{D}|$ images with lowest s_I scores

Algorithm 1: Definition of splits for VG-RMATTERS

Split	Majority Proportion	Average Entropy
VG-IMP [27]	0.62	0.55
VG-RMATTERS	0.44	0.68

Table 1: Entropy and proportion of the majority relation in VG-IMP [27] and VG-RMATTERS for the top 50 pairs of object categories (81% of examples). VG-RMATTERS shows a greater diversity of relations.

5.3. Evaluating Relation Diversity

Lu *et al*. [15] observe that precision metrics are not well adapted to VRD evaluation, because relation annotations are incomplete. Thus measuring the precision and F1 scores risks penalizing predictions of true relations that have not been annotated. Hence, following most recent works, we measure *recall@k*, which corresponds to the fraction of ground truth annotations in top *k* confident relationship predictions. Most recent works focus on the model performance on the image macro recall:

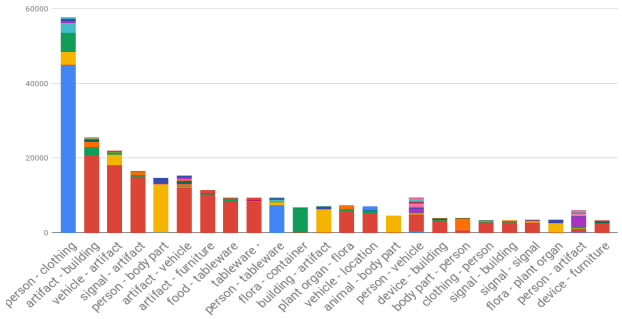
$$\text{IMMACRO } R@k = \frac{1}{|D|} \sum_{I \in D} R@k_I \tag{12}$$

where $R@k_I$ is the recall for image I.

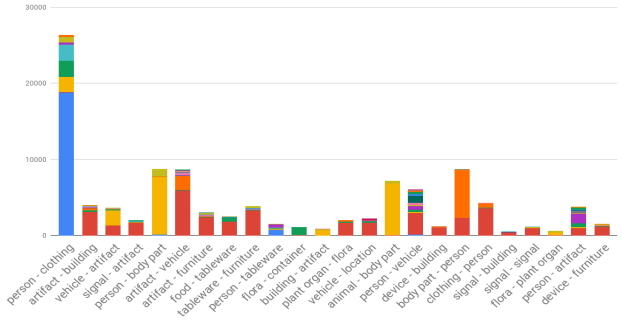
However, retrieving uncommon relations is necessary for many applications where important classes have a low number of training examples. To complement this evaluation, we propose to also use class macro recall, which averages the recall over each relation class:

$$\text{CLSMACRO } R@k = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} R@k_r \tag{13}$$

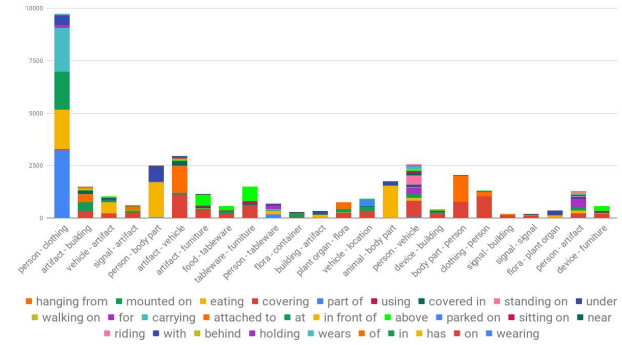
where $R@k_r$ is the recall of relations of class r. The underlying intuition here, is that by averaging on every example, the image macro recall exhibits the ability of the network to reliably retrieve most relations. We focus the retrieval of



(a) Relation train distribution in VG-IMP [27]



(b) Relation test distribution in VG-IMP [27]



(c) Relation test distribution in VG-RMATTERS (ours)

Figure 4: Distribution of relations for each object category on VG-IMP [27] and VG-RMATTERS.

most relations but also of relations that occur more infrequently and differentiate images from one another.

Finally, we do not constrain the predictions to one relation per object pair, as several relations may be true for each object pair (reported as K=50 in other works).

6. Experiments

We evaluate our model on the Visual Genome dataset and compare it with state of the art approaches.

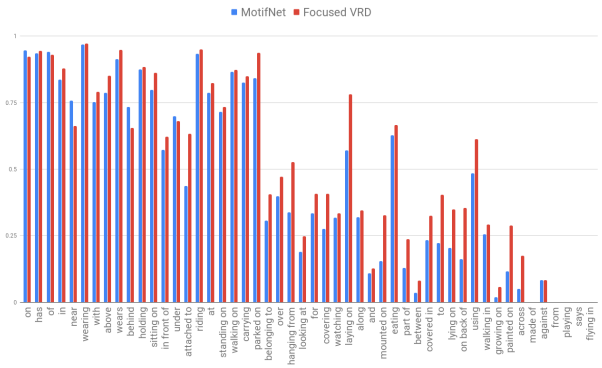


Figure 5: Recall per relation class for MOTIFNET [32] and FOCUSEDVRD (Ours) on VG-RMATTERS, ordered by decreasing frequency in the training set. Focusing predictions on important object pairs, FOCUSEDVRD is able to predict a more diverse set of relations and to increase the recall of rarer relations while keeping a competitive global recall.

6.1. Experimental settings

Datasets **Visual Genome** [12] has 108,077 images. The split proposed in [27] is restricted to 150 object classes and 50 relations with an average of 22 relationships annotations per image. 75% of images are used for training, 5% for validation and 20% for test. The VG-RMATTERS. split has the same classes with a partitioning of 85%/5%/10%.

Implementation details We use the same training procedure as [32]: the combined loss is the sum of cross-entropy losses for object and relation classifications. As in [32], it is minimized by SGD with momentum, with a learning rate of 6.10^{-3} and 6 images per batch.

Evaluation tasks The performance of several methods are compared on two tasks. **Relation detection** (RELCLS): given ground truth object bounding boxes and classes, the task consists in predicting true relations between any pair of objects. We also evaluate models on the **scene graph classification** (SGCLS), where the task is to predict object and relation classes.

6.2. Comparative results

In Table 2, we compare our method to state of the art approaches on VG-IMP [27] and VG-RMATTERS. PIXEL2GRAPH [17] iteratively refines object and relationship heatmaps with a stacked hourglass network making use of global context. IMP [27] refines relation and object representations by passing mesages through the scene graph. MOTIFNET [32] captures higher order correlations between objects and relationships using LSTM layers. SGP [7] is

	SGCLS			RELCLS					
	IMMACRO			IMMACRO			CLSMACRO		
	R@20	R@50	R@100	R@20	R@50	R@100	R@20	R@50	R@100
FREQ BASELINE	31.0	39.2	43.9	52.9	69.8	80.0	11.2	22.1	33.7
PIXEL2GRAPH [17]	-	26.5	30.0	-	68.0	75.2	-	-	-
IMP [27]	-	43.4	47.2	-	75.2	83.6	-	-	-
SGP [7]	-	**45.5**	**50.8**	-	80.8	88.2	-	-	-
MOTIFNET [32]	37.6	44.5	47.7	66.6	81.2	88.3	15.8	27.7	38.1
BASELINE (Ours)	37.1	44.1	47.2	**66.7**	**81.4**	**88.7**	17.6	30.2	41.3
FOCUSEDVRD (Ours)	36.8	43.8	47.0	66.6	81.0	87.7	**18.8**	**32.3**	**44.4**

Table 2: Results on VG-IMP. Recalls are in % and evaluated without scene graph constraints (k=50). The relevance factor increases the CLSMACRO recall at all sizes of scene graphs but slightly decreases the IMMACRO.

	SGCLS			RELCLS						N pairs
	IMMACRO			IMMACRO			CLSMACRO			
	R@10	R@20	R@100	R@10	R@20	R@100	R@10	R@20	R@100	Pairs@10
MOTIFNET [32]	26.3	38.0	55.5	39.6	58.4	87.8	11.8	19.8	46.6	8.6
	26.7	38.5	56.3	40.3	58.9	**88.3**	12.6	21.1	47.8	8.6
RC	27.8	39.8	**56.4**	**44.2**	**62.3**	87.7	**14.5**	**24.0**	**52.6**	7.4
BP	28.9	40.3	56.2	43.2	61.4	87.9	14.2	23.4	51.7	7.8
RC + BP	**29.4**	**41.0**	**56.4**	44.0	**62.3**	**88.3**	14.3	23.9	52.4	7.6

Table 3: Ablation study on VG-RMATTERS. Recalls are in % and evaluated without scene graph constraints (k=50). Ablation study on VG-RMATTERS. In a more balanced dataset, the relevance factor increases both CLSMACRO and IMMACRO recalls, especially for smaller scene graphs.

a permutation invariant graph predictor that refines predictions from MOTIFNET using attention over linguistic and visual features of neighbors. RELDN [34] is not included because we noticed that the implementation resulted in a different evaluation protocol, where object pairs without relations are filtered out. Since we aim to improve the precision of VRD and to output more relevant relations, we will focus on results on small scene graphs.

On VG-IMP [27], our baseline slightly outperforms MOTIFNET. FOCUSEDVRD provides a significant improvement in the class macro recall, from 38.1% to 44.4% and a 3 points improvement over our baseline. This does not translate into the image macro recall, which slightly drops from 88.7% to 87.7% for the R@100 but is competitive on small graphs. Scores on smaller scene graphs show that when higher precision is needed, our model is competitive. Furthermore, the higher macro recall shows that our model performs much better on detecting less common relations and is thus able to highlight information that makes an image stand out. This translates into a higher recall when relations are more balanced for each object pair, on VG-RMATTERS, where the R@10 image macro recall increases from 39.6% for MOTIFNET to 44.0% for FOCUSEDNET.

6.3. Ablation study

Table 3 shows the results of our model on VG-RMATTERS and the influence of Relevance Modeling, with a Relevance Classifier (RC) and Binary Potential (BP). With the relevance factor, class and image macro recalls of smaller scene graphs increase. RC has a higher recall than BP for most settings, due to a lower number of predicted pairs. BP is useful at a higher count of relations, smoothing the relation scores and increasing the number of selected object pairs. The combination of both factors acts as ensembling by improving this selection. Finally, a higher class macro recall is linked with improved performance on R@10 and R@20 but does does not necessarily lead to improved R@100.

Figure 6 shows the the output FOCUSEDVRD with RC and BP compared to the baseline and the ground truth. The model is able to detect and focus on important object pairs. This removes the false relation (*flower*, *on*, *chair*) but in-

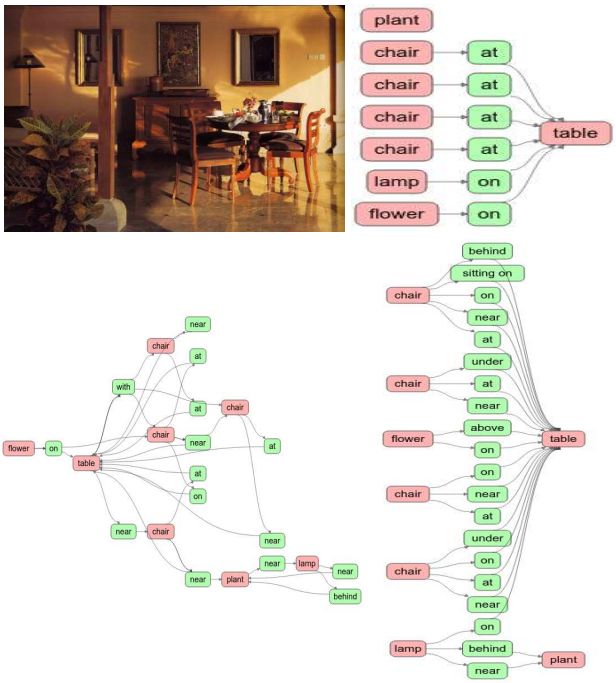


Figure 6: Image example from VG and associated scene graph (*top right*), with extracted scene graphs of our baseline (bottom left) and FOCUSEDVRD (RC+BP) (*bottom right*).

creases the risk of predicting false relations, such as (*chair, sitting on, table*).

6.4. Discussion

Predicting the relevance of a relation is difficult due to the various factors it depends on, such as the size, location, distances of objects as well as who is annotating. Indeed, only 47% of the descripted regions of VG [12] are annotated with a relation, which suggests that the relations deemed relevant to a human being can be very different to another human being and makes the signal during training very noisy. Hence our choice to augment the learned classifier with frequency priors. Results in Tables 2 and 3 show that our method significantly improves the class macro recall. Furthermore, in Figure 5, we observe that by focusing on a smaller number of pairs, FOCUSEDVRD outperforms MOTIFNET on most relations, with a significant margin for several rare relations, such as *laying on* and *painted on*. These improvements translate into significant improvement of the image macro recall in the case of VG-RMATTERS, which was constructed so that predicting the most frequent relation is less viable. However, overall performance is comparable between both datasets. This might come from the higher proportion of 'person - clothing' pairs. It makes the task of finding which pairs to annotate easier and reduces the overall frequency of other relations. yu

7. Conclusion and Perspectives

In order to tackle the subject of Visual Relation Detection (VRD), we proposed a novel scene graph generation process integrating an approximation of the relative relevance of relations. This enables it to better handle the diversity of relations than MOTIFNET, a state of the art method for Relation Classification. The gain in performance is most relevant on small scene graphs, which we focus on to decrease output noise.

A deep neural network is trained to predict the relevance of each pair of objects in an image. The learnt relevance score is added to a statistics-based prior, measured on the training set. Integrated to the scene graph generation process, this new factor increases the focus of predictions on small sets of relevant pairs. We showed that this improves relation detection in the case of limited training data, increasing the average recall of classes from 38.1% to 44.4%. Thus, this method is most useful for datasets where, like Visual Genome, relations are highly skewed towards a small set of relations and when downstream applications rely on detecting anomalies rather than capturing a high number of relations. In such cases, it is important to predict several probable relations on a small number of object pairs, so that more meaningful relations are extracted. To further show the impact of this approach, we proposed a new split of the Visual Genome dataset [12] with a more balanced relation distribution in the test set. This was motivated by the small impact that performance on uncommon relations has on usual performance metrics.

Finally, we showed that both elements of the relevance classifier should be used in different applications. For applications requiring a high recall, the added priors are beneficial. However, when focusing on a high precision, one should only use the relevance classifier, which focuses predictions on a smaller number of relations.

Perspectives Study of the model output shows that weighing by relevance increases the risk of predicting false positives. Thresholding probabilities in order to remove less likely relations would help increase the precision of the model. Furthermore, on heavily skewed test sets, methods aimed at targeting data imbalance are not very effective, being too heavily punished by the most frequent classes. We plan to explore these methods, such as under-sampling, in order to improve relation classification on VG-RMATTERS. Finally, the proposed split is limited by the relation distribution in Visual Genome. For this reason we would like to collect complementary images containing rare relations and with a higher diversity of object pairs.

References

[1] A. C. Berg, T. L. Berg, H. Daume, J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, A. Sood, K. Stratos, and K. Yamaguchi. Understanding and Predicting Importance in Images. In *CVPR*, 2012.

[2] Y. W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng. HICO: A benchmark for recognizing human-object interactions in images. In *ICCV*, 2015.

[3] R. Girshick. Fast R-CNN. In *ICCV*, 2015.

[4] G. Gkioxari, R. Girshick, P. Dollár, and K. He. Detecting and Recognizing Human-Object Interactions. In *CVPR*, 2018.

[5] Y. Goyal, T. Khot, A. Agrawal, D. Summers-Stay, D. Batra, and D. Parikh. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. *International Journal of Computer Vision*, 2019.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016.

[7] R. Herzig, M. Raboh, G. Chechik, J. Berant, and A. Globerson. Mapping Images to Scene Graphs with Permutation-Invariant Structured Prediction. In *NIPS*, 2018.

[8] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural computation*, 1997.

[9] J. Johnson, A. Gupta, and L. Fei-Fei. Image Generation from Scene Graphs. In *CVPR*, 2018.

[10] J. Johnson, R. Krishna, M. Stark, L.-j. Li, D. A. Shamma, M. S. Bernstein, L. Fei-fei, and Y. Labs. Image Retrieval using Scene Graphs. In *CVPR*, 2015.

[11] A. Karpathy and L. Fei-Fei. Deep Visual-Semantic Alignments for Generating Image Descriptions. *Transactions on Pattern Analysis and Machine Intelligence*, 39(4):664–676, 2015.

[12] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. 2016.

[13] Y. Li, W. Ouyang, X. Wang, and X. Tang. ViP-CNN: Visual Phrase Guided Convolutional Neural Network. In *CVPR*, 2017.

[14] K. Liang, Y. Guo, H. Chang, and X. Chen. Visual Relationship Detection with Deep Structural Ranking. In *AAAI*, 2018.

[15] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detection with language priors. In *ECCV*, 2016.

[16] I. Misra, C. L. Zitnick, M. Mitchell, and R. Girshick. Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. In *CVPR*, 2016.

[17] A. Newell and J. Deng. Pixels to Graphs by Associative Embedding. In *NIPS*, 2017.

[18] J. Peyre, I. Laptev, C. Schmid, and J. Sivic. Weakly-supervised learning of visual relations. In *ICCV*, 2017.

[19] F. Plesse, A. Ginsca, B. Delezoide, and F. Prêteux. Visual Relationship Detection Based on Guided Proposals and Semantic Knowledge Distillation. In *ICME*, 2018.

[20] V. Ramanathan, C. Li, J. Deng, W. Han, Z. Li, K. Gu, Y. Song, S. Bengio, C. Rossenberg, and F. F. Li. Learning semantic relationships for better action retrieval in images. In *CVPR*, 2015.

[21] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In *CVPR*, 2016.

[22] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NIPS*, 2015.

[23] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*, 2015.

[24] D. Teney, L. Liu, and A. Van Den Hengel. Graph-Structured Representations for Visual Question Answering. In *CVPR*, 2017.

[25] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015.

[26] S. Woo, D. Kim, K. Daejeon, D. E. Cho, and I. E. So Kweon. LinkNet: Relational Embedding for Scene Graph. In *NIPS*, 2018.

[27] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei. Scene Graph Generation by Iterative Message Passing. In *CVPR*, 2017.

[28] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *ICML*, 2015.

[29] G. Yin, L. Sheng, B. Liu, N. Yu, X. Wang, J. Shao, and C. C. Loy. Zoom-Net: Mining Deep Feature Interactions for Visual Relationship Recognition. In *ECCV*, 2018.

[30] R. Yu, A. Li, V. I. Morariu, and L. S. Davis. Visual Relationship Detection With Internal and External Linguistic Knowledge Distillation. In *ICCV*, 2017.

[31] R. Zellers, Y. Bisk, A. Farhadi, Y. Choi, and P. G. Allen. From Recognition to Cognition: Visual Commonsense Reasoning. In *CVPR*, 2019.

[32] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi. Neural Motifs: Scene Graph Parsing with Global Context. In *CVPR*, 2018.

[33] J. Zhang, K. J. Shih, A. Elgammal, A. Tao, and B. Catanzaro. Graphical Contrastive Losses for Scene Graph Parsing. In *CVPR*, 2019.

[34] R. Zhang, L. Lin, G. Wang, M. Wang, and W. Zuo. Hierarchical Scene Parsing by Weakly Supervised Learning with Image Descriptions. *Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[35] Y. Zhu, S. Jiang, and X. Li. Visual relationship detection with object spatial distribution. In *ICME*, 2017.