This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Dense Relational Captioning: Triple-Stream Networks for Relationship-Based Captioning

Dong-Jin Kim[1] Jinsoo Choi[1] Tae-Hyun Oh [2] In So Kweon[1]
[1]KAIST, South Korea. [2] MIT CSAIL, Cambridge, MA.
[1]{djnjusa,jinsc37,iskweon77}@kaist.ac.kr [2]taehyun@csail.mit.edu

Abstract

Our goal in this work is to train an image captioning model that generates more dense and informative captions. We introduce "relational captioning," a novel image captioning task which aims to generate multiple captions with respect to relational information between objects in an image. Relational captioning is a framework that is advantageous in both diversity and amount of information, leading to image understanding based on relationships. Part-of-speech (POS, i.e. *subject-object-predicate categories) tags can be assigned to every English word. We leverage the POS as a prior to guide the correct sequence of words in a caption. To this end, we propose a multi-task triple-stream network (MTTSNet) which consists of three recurrent units for the respective POS and jointly performs POS prediction and captioning. We demonstrate more diverse and richer representations generated by the proposed model against several baselines and competing methods.*

1. Introduction

Human visual system has the capability to effectively and instantly collect a holistic understanding of contextual associations among objects in a scene [16, 23] by densely and adaptively skimming the visual scene through the eyes, *i.e*. the saccadic movements. Such instantly extracted rich and dense information allows humans to have the superior capability of object-centric visual understanding. Motivated by this, in this work, we present a new concept of scene understanding, called *dense relational captioning*, that provides dense but selective, expressive, and relational representation in a human interpretable way, i.e., via captions.

Richer representation of an image often leads to numerous potential applications or performance improvements of subsequent computer vision algorithms [22, 23]. In order to achieve richer object-centric understanding, Johnson *et al*. [12] proposed a framework called DenseCap that generates captions for each of the densely sampled local image regions. These regional descriptions facilitate both rich and

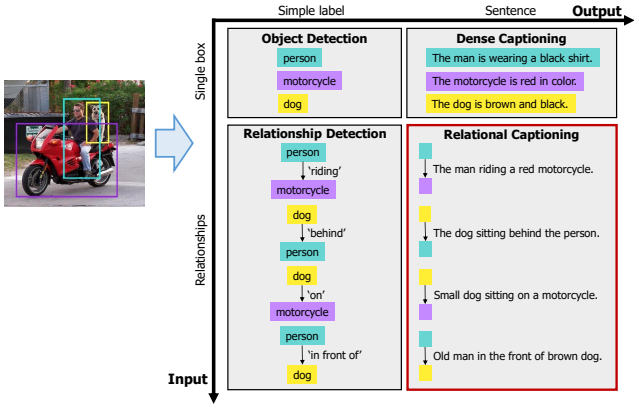

Figure 1: Overall description of the proposed relational captioning framework. Compared to traditional frameworks, our framework is advantageous in both interaction understanding and high-level interpretation.

dense semantic understanding of a scene in a form of interpretable language. However, the information in the image that we want to acquire includes not only the information of the object itself but also the *interaction* with other objects or the environment.

As an alternative way of representing an image, we focus on dense *relationships* between objects. In the context of human cognition, there has been a general consensus that objects and particular environments near the target object affect search and recognition efficiency. Understanding the relationships between objects clearly reveal object interactions and object-attribute combinations [11, 14, 20].

Interestingly, we observe that the annotations done by humans on computer vision datasets predominantly contain relational forms; in Visual Genome [15] and MS COCO [19] caption datasets, most of the labels take the format of subject-predicate-object more so than subject-predicate. Moreover, UCF101 [31] action recognition dataset contains 85 actions out of 101 (84.2%) that are described in terms of human interactions with other objects or surroundings. These aspects tell us that understanding interaction and relationships between objects facilitate a major

component in visual understanding of object-centric events.

In this regard, we introduce a novel captioning framework *relational captioning* that can provide diverse and dense representations from an image. In this task, we first exploit the relational context between two objects as a representation unit. This allows generating a combinatorial number of localized regional information. Secondly, we make use of captioning and its ability to express significantly richer concepts beyond the limited label space of object classes used in object detection tasks. Due to these aspects, our relational captioning expands the regime further along the label space both in terms of density and complexity, and provides richer representation for an image.

Our main contributions are summarized as follows. (1) We introduce *relational captioning*, a new captioning task that generates captions with respect to relational information between objects in an image. (2) In order to efficiently train the relational caption information, we propose the *multi-task triple-stream network* (MTTSNet) that consists of three recurrent units trained via multi-task learning. (3) We show that our proposed method is able to generate denser and more diverse captions by evaluating on our relational captioning dataset augmented from Visual Genome (VG) [15] dataset. (4) We introduce several applications of our framework, including "caption graph" generation which contains richer and more diverse information than conventional scene graphs.

2. Related Work

Our work relates to two topics: image captioning and relationship detection. In this section, we categorize and review related work on these topics.

Image captioning. By virtue of deep learning and the use of recurrent neural network (*e.g*. LSTM [9]) based decoders, image captioning [24] techniques have been extensively explored [1, 7, 10, 13, 21, 28, 33, 37, 39, 41]. One of the research issues in captioning is the generation of diverse and informative captions. Thus, learning to generated diverse captions has been extensively studied recently [2, 4, 5, 29, 32, 34]. As one of the solutions, the dense captioning (DenseCap) task [12] was proposed which uses diverse region proposals to generate localized descriptions, extending the conventional holistic image captioning to diverse captioning that can describe local contexts. Moreover, our *relational captioning* is able to generate even more diverse caption proposals than dense captioning by considering *relations* between objects.

Yang *et al*. [38] improves the DenseCap model by incorporating a global image feature as context cue as well as a region feature of the desired objects with a late fusion. Motivated by this, in order to implicitly learn dependencies of subject, object and union representations, we incorporate a triple-stream LSTM for our captioning module.

Visual relationship detection (VRD). Understanding visual relationships between objects have been an important concept in various tasks. Conventional VRD usually deals with predicting the subject-predicate-object (in short, `subj-pred-obj`). A pioneering work by Lu *et al*. [20] formalizes the VRD task and provides a dataset, while addressing the subject (or object) and predicate classification models separately. On the other hand, similar to VRD task, scene graph generation (a task to generate a structured graph that contains the context of a scene) has also started to be explored [18, 35, 36, 43].

Although the VRD dataset is larger (100 object classes and 70 predicates) than Visual Phrases, it is still inadequate to handle the real world scale. The Visual Genome (VG) dataset [15] for relationship detection consists of 31k predicate types and 64k object types giving the number of possible combinations of relationship triplets too diverse for the state-of-the-art VRD based models. This is because the labels consist of the various combinations of words (*e.g*. 'little boy,' 'small boy,' *etc*.) As a result, only the simplified version of VG relationship dataset has been studied. On the contrary, our method is able to generate relational captions by tokenizing the whole relational expressions into words, and learning from them.

While the recent state-of-the-art VRD [17, 20, 26, 42, 40] or scene graph generation works [18, 35, 36, 43] mostly use language priors to detect relationships, we directly learn the relationship as a descriptive language model. In addition, the expressions of traditional scene graph generation or VRD task are restricted to `subj-pred-obj` triplets, whereas the relational captioning is able to provide additional information such as attributes or noun modifiers by adopting free-form natural language expressions.

In summary, dense captioning facilitates a natural language interpretation of regions in an image, while VRD can obtain relational information between objects. Our work combines both axes, resulting in much denser and diverse captions than DenseCap. That is, given B region proposals in an image, we can obtain $B(B-1)$ relational captions, whereas DenseCap returns only B captions.

3. Multi-task Triple-Stream Networks

Our relational captioning is defined as follows. Given an input image, a bounding box detector generates various object proposals and a captioning module predicts combinatorial captions with POS labels describing each pair of objects. Figure 2 shows the overall framework of the proposed relational captioning model, which is mainly composed of a localization module based on the region proposal network (RPN) [27], and a triple-stream RNN (LSTM [9]) module for captioning. Our network supports end-to-end training with a single optimization step that allows joint localization, combination, and description with natural language.

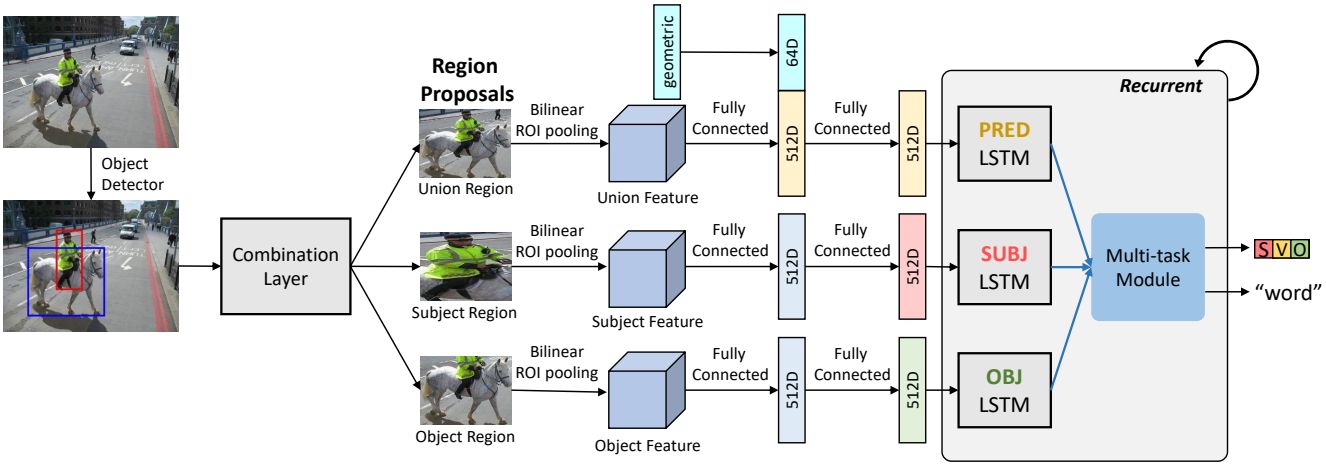

Figure 2: Overall architecture of the proposed multi-task triple-stream networks. Three region features (Union, Subject, Object) come from the same shared branch (Region Proposal Network), and for *subject* and *object* features, the first intermediate FC layer share weights (depicted in the same color).

Given an image, RPN generates object proposals. Then, the combination layer takes a pair consisting of a *subject* and an *object* at a time. To take the surrounding context information into account, we utilize the *union* region of the *subject* and *object* regions, in a way similar to using the global image region as side information by Yang *et al*. [38]. This feature of triplets (*subject*, *object*, *union*) are fed to the triple-stream LSTMs, where each stream takes its own purpose, *i.e*. *subject*, *object*, and *union*. Given this triplet feature, the triple-stream LSTMs collaboratively generate a caption and POS classes of each word. We describe these processes as follows.

3.1. Region Proposal Networks

Our network uses fully convolutional layers of VGG-16 [30] up to the final pooling layer (*i.e*. `conv5_3`) for extracting the spatial features via the bilinear ROI pooling [12]. The object proposals are generated by localization layers. It takes the feature tensor, and proposes B regions (user parameter) of interest. Each proposed region has its confidence score, region feature of shape $512\times7\times7$, and coordinates $b=(x, y, w, h)$ of the bounding box with center (x, y), width w and height h. We process it into vectorized features (of shape $D=512$) using two fully-connected (FC) layers. This encodes the appearance of each region into a feature, called *region code*. Once the region codes are extracted, they are reused for the following processes.

To generate relational proposals, we build pairwise combinations of B region proposals, where in turn we get $B(B-1)$ possible region pairs. We call this layer the *combination layer*. A distinctive point of our model with the previous dense captioning works [12, 38], is that while the works regard each region proposal as an independent target to describe and produce B number of captions, we consider their pairwise combinations $B(B-1)$, which are much denser and explicitly expressible in term of relationships. Also, we can asymmetrically use each entry of a pair by assigning the roles of the regions, *i.e*. (*subject*, *object*).

Furthermore, motivated by Yang *et al*., where the global context of an image improves the captioning performance, we leverage an additional region, the *union* region b_u of (*subject*, *object*). In addition, to provide relative spatial information, we append a geometric feature for the *subject* and *object* box pair, *i.e*. (b_s, b_o) to the *union* feature before the FC layers. Given two bounding boxes b_s and b_o, the geometric feature $\mathbf{r}$ is defined similarly to [25] as

$$\mathbf{r} = \left[\frac{x_o - x_s}{\sqrt{w_s h_s}}, \frac{y_o - y_s}{\sqrt{w_s h_s}}, \sqrt{\frac{w_o h_o}{w_s h_s}}, \frac{w_s}{h_s}, \frac{w_o}{h_o}, \frac{b_s \bigcap b_o}{b_s \bigcup b_o} \right] \in \mathbb{R}^6. \quad (1)$$

By concatenating the *union* feature with $\mathbf{r}$ which is passed through an additional FC layer, the shape of this feature is $D+64$. Then, the dimension of the *union* region code is reduced by the following FC layers. This stream of operations is illustrated in Fig. 2. The three features extracted from the *subject*, *object*, and *union* regions are fed to each LSTM described in the following sections.

3.2. Relational Captioning Networks

Relational caption generation takes the relational information of the object pairs into account. However, expressing the relationship in a sentence has been barely studied. Therefore, we design a new network that deals with relational captions, called the *multi-task triple-stream network*.

From the region proposal network, a triplet of region codes are fed as input to LSTM cells, so that a sequence of words (caption) is generated. In the proposed relational region proposal, a distinctive facet is to provide a triplet of region codes consisting of *subject*, *object*, and *union* regions, which virtually corresponds to the POS of a sentence (`subj-pred-obj`). This correspondence between regions

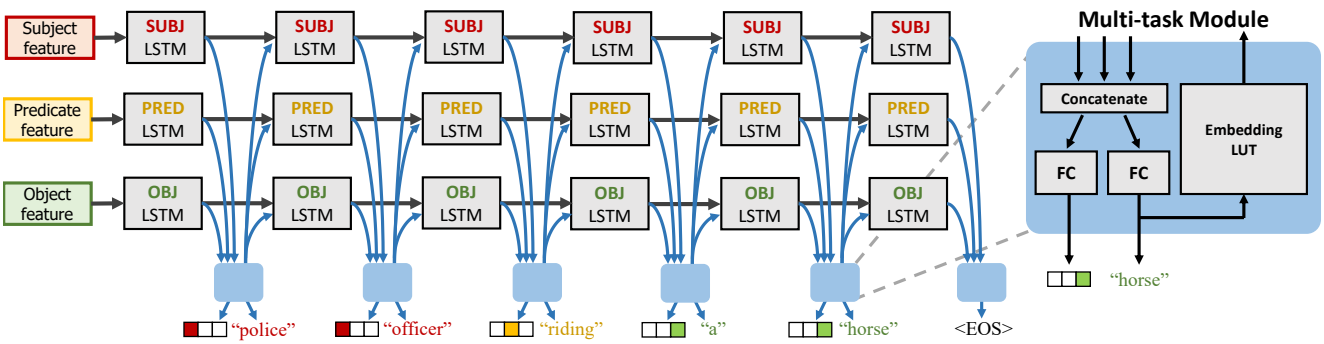

Figure 3: An illustration of the unrolled triple-stream LSTM. Our model consists of two major parts: triple-stream LSTM and a multi-task module. The multi-task module jointly predicts a caption word and its POS class (`subj-pred-obj`, illustrated as three cells colored according to the POS class), as well as the input vector for the next time step.

in a triplet and POS information leads to the following advantages: 1) input features can be adaptively merged depending on its POS and fed to the caption generation module, and 2) the POS prior on predicting a word can be effectively applied to caption generation. However, leveraging and processing these input cues are non-trivial.

For the first advantage, in order to derive POS aware inference, we propose *triple-stream networks*, which are three separate LSTMs respectively corresponding to `subj-pred-obj`. The outcomes of LSTMs are combined via concatenation. For the second advantage, during a word prediction, we jointly infer its POS class via multi-task inference. This POS class prediction acts as a prior for the word prediction of a caption during the learning phase.

Triple-Stream LSTMs. Intuitively, the region codes of *subject* and *object* would be closely related to the subject and object related words in a caption, while the *union* and geometric features may contribute to the predicate. In our relational captioning framework, the LSTM modules must adaptively take input features into account according to which POS decoding stage it is for a caption.

As shown in Fig. 2, the proposed triple-stream LSTM module consists of three separate LSTMs, each of which is in charge of the *subject*, *object* and *union* region codes respectively. At each step, the triple-stream LSTMs generate three embedded representations separately, and a single word is predicted by consolidating the three processed representations. The embedding of the predicted word is distributed into all three LSTMs as inputs and is used to run the next step in a recursive manner. Thus in each step, each entry of the triplet input is used differently, which allows more flexibility than a single LSTM as used in traditional captioning models [12, 33]. In other words, the weights of the input cue features change at every recursive step according to which POS the word being generated belongs to.

Multi-task with POS Classification. On top of this concatenation, we utilize the POS information to more effectively train the relational captioning model. Relational captioning generates a sequence of words in `subj-pred-obj` order, *i.e.* the order of POS. For each word prediction, in a *multi-task* module in Fig. 3, we also classify the POS class of the predicted word, so that it encourages the caption generation to follow the word order in the POS order.

When three representations for each POS are to be consolidated, one option can be to consolidate them in an early step, called *early fusion*. This results in a single LSTM with the fusion of the three region codes (*e.g.* concatenation of three codes). However, as reported by Yang *et al.* [38], this early fusion approach also shows lower performance than that of late fusion methods. In this regard, we adopt a *late fusion* for a multi-task module. The layer basically concatenates the representation outputs from the triple-stream LSTMs, but due to the recurrent multi-task modules, it is able to generate sophisticated representations.

We empirically observe that this multi-task learning with POS helps not only the shared representation to become richer but also guides the word predictions, and thus helps to improve the captioning performance overall. We hypothesize that the POS task provides distinctive information that may help learn proper representations from the triple-stream LSTMs. Since each POS class prediction tightly relies on respective representations from each LSTM, *e.g. pred*-LSTM closely related to `pred` of POS, the gradients generated from the POS classification would be back-propagated through the indices of the concatenated representation according to the class. By virtue of this, the multi-task triple-stream LSTMs are able to learn the representation in such a way that it can predict plausible words for each time step. Therefore, our model can generate appropriate words according to the POS at a given time step.

Loss functions. Training our relational captioning model can be mainly divided into captioning loss and detection loss. Specifically, the proposed model is trained to minimize the following loss function:

$$\mathcal{L} = \mathcal{L}_{cap} + \alpha\mathcal{L}_{POS} + \beta\mathcal{L}_{det} + \gamma\mathcal{L}_{box}, \qquad (2)$$

where $\mathcal{L}_{cap}$, $\mathcal{L}_{POS}$, $\mathcal{L}_{det}$, and $\mathcal{L}_{box}$ denote captioning loss, POS classification loss, detection loss, and bounding box regression loss respectively. α, β, and γ are the balance parameters (we set them to 0.1 for all experiments).

The first two terms are for captioning and the next two terms are for the region proposal. $\mathcal{L}_{cap}$ and $\mathcal{L}_{POS}$ are cross-entropy losses at every time step for each word and POS classification respectively. For each time step, $\mathcal{L}_{POS}$ measures a 3-class cross entropy loss. $\mathcal{L}_{det}$ is a binary logistic loss for foreground/background regions, while $\mathcal{L}_{box}$ is a smoothed L1 loss [27].

4. Experiments

In this section, we provide the experimental setups, competing methods and performance evaluation of relational captioning with both quantitative and qualitative results.

4.1. Relational Captioning Dataset

Since there is no existing dataset for the relational captioning task, we construct a dataset by utilizing VG relationship dataset version 1.2 [15] which consists of 85200 images with 75456/4871/4873 splits for train/validation/test sets respectively. We tokenize the relational expressions to form natural language expressions, and for each word, we assign the POS class from the triplet association.

However, VG relationship datasets show limited diversity in the words used. Therefore, by only using relational expressions to construct data, the captions generated from a model tends to be simple (*e.g.* "building-has-window"). Even though our model may enable richer concepts and expressions, if the training data does not contain such concepts and expressions, there is no way to actually *see* this. In order to validate the diversity of our relational captioner, we need to make our relational captioning dataset to have more natural sentences with rich expressions.

Through observation, we noticed that the relationship dataset labels lack *attributes* describing the subject and object, which are perhaps what enriches the sentences the most. Therefore, we utilize the *attribute labels* of VG data to augment existing relationship expressions. More specifically, we simply find the attribute that matches the subject/object of the relationship label and attach it to the subj/obj caption label. In particular, if an attribute label describes the same subject/object for a relationship label while associated bounding box overlaps enough, the label is considered to be *matched* to the subject/object in the relationship label. After this process, we obtain 15595 vocabularies for our relational captioning dataset (11447 vocabularies before this process). We train our caption model with this data, and report its result in this section. In addition, we provide a holistic image captioning performance and various analysis such as comparison with scene graph generation.

	mAP (%)	Img-Lv. Recall	METEOR
`Direct Union`	–	17.32	11.02
`Union`	0.57	25.61	12.28
`Union+Coord.`	0.56	27.14	13.71
`Subj+Obj`	0.51	28.53	13.32
`Subj+Obj+Coord.`	0.57	30.53	14.85
`Subj+Obj+Union`	0.59	30.48	15.21
`TSNet (Ours)`	**0.61**	**32.36**	**16.09**
`Union (w/MTL)`	0.61	26.97	12.75
`Subj+Obj+Coord (w/MTL)`	0.63	31.15	15.31
`Subj+Obj+Union (w/MTL)`	0.64	31.63	16.63
`MTTSNet (Ours)`	**0.88**	**34.27**	**18.73**
`Neural Motifs [43]`	0.25	29.90	15.34

Table 1: Ablation study for relational dense captioning task on relational captioning dataset.

4.2. Relational Dense Captioning: Ablation Study

Baselines. Since no direct work for relational captioning exists, we implement several baselines by modifying the most relevant methods, which facilitate our ablation study.

• `Direct Union` has the same architecture with *DenseCap* [12], but of which RPN is trained to directly predict union regions. The union region is used to generate captions by one LSTM.

• `Union` also resembles *DenseCap* [12] and `Direct union`, but its RPN predicts individual object regions. The object regions are paired as (subject, object), and then a union region from each pair is fed to a single LSTM for captioning. Also, we implement two additional variants: `Union (w/MTL)` additionally predicts the POS classification task, and `Union+Coord.` appends the geometric feature to the region code of the union.

• `Subj+Obj` and `Subj+Obj+Union` models use the concatenated region features of (subject, object) and (subject, object, union) respectively and pass them through a single LSTM (early fusion approach). Also, `Subj+Obj+Coord.` uses the geometric feature instead of the region code of the union. Moreover, we evaluate the baselines, `Subj+Obj+{Union,Coord}` with POS classification (MTL loss).

• `TSNet` denotes the proposed triple-stream LSTM based model without a branch for POS classifier. Each stream takes the region codes of (subject, object, union + coord.) separately. `MTTSNet` denotes our final model, multi-task triple-stream network with POS classifier.

Evaluation metrics. Motivated by the evaluation metric suggested for dense captioning task [12], we suggest a new evaluation metric for relational dense captioning. We report the mean Average Precision (mAP) which measures both localization and language accuracy. As suggested by Johnson *et al.*, we use METEOR score [6] with thresholds $\{0, 0.05, 0.10.15, 0.2, 0.25\}$ for language, and IOU thresholds $\{0.2, 0.3, 0.4, 0.5, 0.6\}$ for localization. The AP values

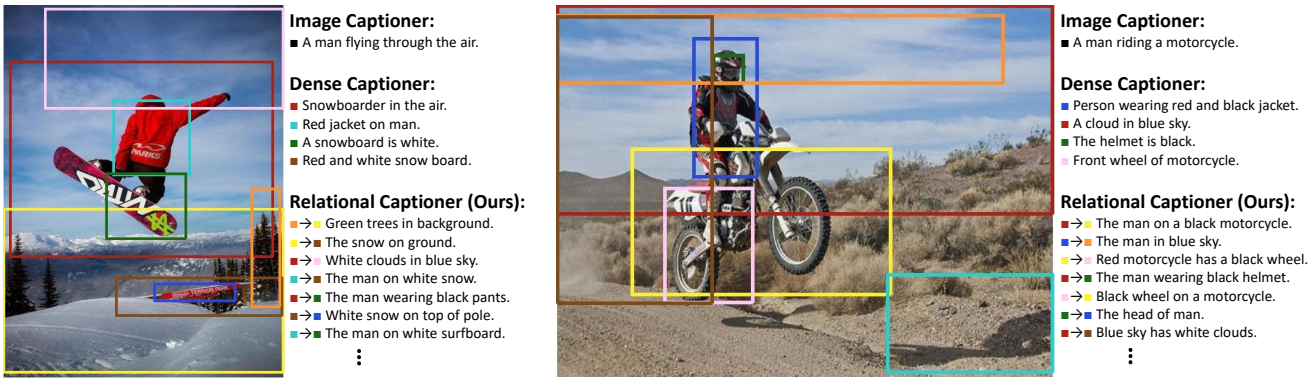

Figure 4: Example captions and region generated by the proposed model. We compare our result with the image captioner [33] and the dense captioner [12] in order to contrast the amount of information and diversity.

	Recall	METEOR	#Caption	Caption/Box
Image Captioner (Show&Tell) [33]	23.55	8.66	1	1
Image Captioner (SCST) [28]	24.04	14.00	1	1
Dense Captioner (DenseCap) [12]	42.63	19.57	9.16	1
Relational Captioner (`Union`)	38.88	18.22	85.84	9.18
Relational Captioner (**`MTTSNet`**)	**46.78**	**21.87**	**89.32**	**9.36**

Table 2: Comparisons of the holistic level image captioning. We compare the results of the relational captioners with that of two image captioners [28, 33] and a dense captioner [12].

obtained by all the pairwise combinations of language and localization thresholds are averaged to get the final mAP score. The major difference of our metric is that, for the localization AP, we measure for both the subject and object bounding boxes with respective ground truths. In particular, we only consider the samples with IOUs of both the subject and object bounding boxes greater than the localization threshold. For all cases, we use percentage as the unit of metric. In addition, we suggest another metric, called "image-level (Img-Lv.) recall." This measures the caption quality at the holistic image level by considering the bag of all captions generated from an image as a single prediction. Given only the aforementioned language thresholds for METEOR *i.e.* without box IOU threshold, we measure the recall of the predicted captions. The metric evaluates the diversity of the produced representations by the model for a given image. Also, we measure the average METEOR score for predicted captions to evaluate the caption quality.

Results. Table 1 shows the performance of the relational dense captioning task on relational captioning dataset. The second and third row sections (2-7 and 8-11th rows) show the comparison of the baselines with and without POS classification (`w/MTL`). In the last row, we show the performance of the state-of-the-art scene graph generator, Neural Motifs [43]. Due to the different output structure, we compare with Neural Motifs trained with the supervision for relationship detection. Similar to the setup in DenseCap [12], we fix the number of region proposals before NMS to 50 for all methods for a fair comparison.

Among the results in the second row section (2-7th rows) of Table 1, our TSNet shows the best result suggesting that the triple-stream component alone is a sufficiently strong baseline over the others. On top of TSNet, applying the MTL loss (*i.e.*, MTTSNet) improves overall performance, and especially improves mAP, where the detection accuracy seems to be dominantly improved compared to the improvement of the other metrics. This shows that *triple-stream LSTM* is the key module that most leverages the MTL loss across other early fusion approaches (see the third row section of the table). As another factor, we can see from Table 1 that the relative spatial information (`Coord.`) and union feature information (`Union`) improves the results. This is because the union feature itself preserves the spatial information to some extent from the 7×7 grid form of its activation. For `Neural Motifs`, other relational captioner baselines including our `TSNet` and `MTTSNet` perform favorably against `Neural Motifs` in all metrics. This is worth noting because handling free-form language generation which we aim to achieve is more challenging than the simple triplet prediction of scene graph generation.

4.3. Holistic Image Captioning Comparison

We also compare our approach with other image captioning frameworks, *Image Captioner* (Show&Tell [33] and SCST [28]), and *Dense Captioner* (DenseCap [12]), in a holistic image description perspective. In order to measure the performance of *holistic image-level* captioning for dense captioning methods, we use Img-Lv. Recall metric defined in the previous section (Recall). We compare them with two relational dense captioning methods, `Union`

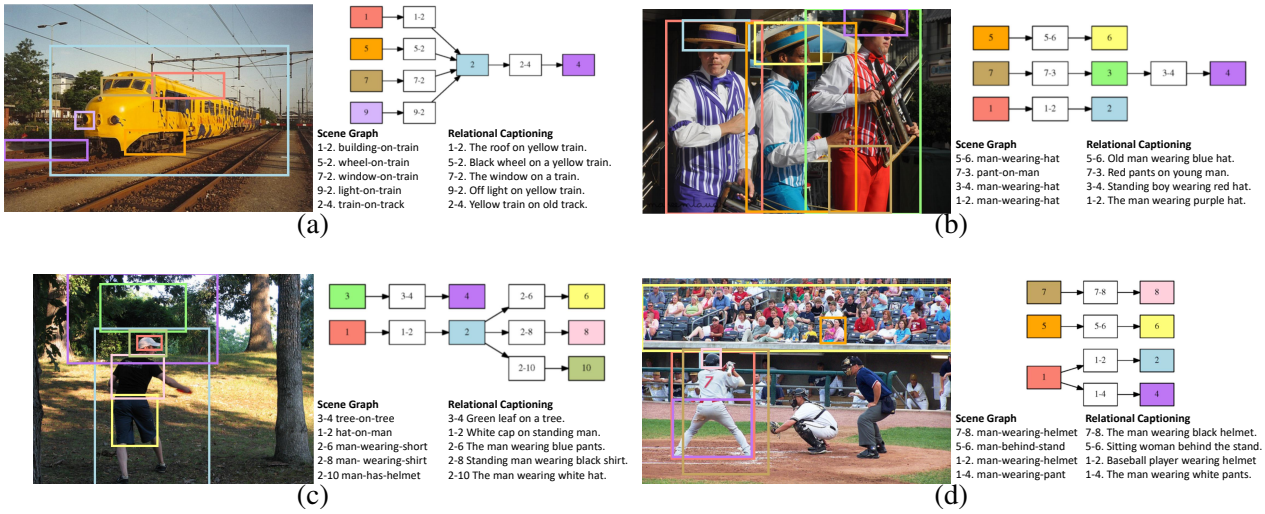

Figure 5: Results of generating "caption graph" from our relational captioniner. In order to compare the diversity of the outputs, we also show the result of the scene graph generator, Neural Motifs [43].

and `MTTSNet`, denoted as *Relational Captioner*. For a fair comparison, for *Dense* and *Relational Captioner*, we adjust the number of region proposals after NMS to be similar, which is different from the setting in the previous section which fixed the number of proposals before NMS.

Table 2 shows the image-level recall, METEOR, and additional quantities for comparison. *#Caption* denotes the average number of captions generated from an input image and *Caption/Box* denotes the average ratio of the number of captions generated and the number of boxes remaining after NMS. Therefore, *Caption/Box* demonstrates how many captions can be generated given the same number of boxes generated after NMS. By virtue of multiple captions per image from multiple boxes, the *Dense Captioner* is able to achieve higher performance than both of the *Image Captioner*s. Compared with the *Dense Captioner*, `MTTSNet` as a *Relational Captioner* can generate an even larger number of captions given the same number of boxes. Hence, as a result of learning to generate diverse captions, the `MTTSNet` achieves higher recall and METEOR. From the performance of `Union`, we can see that it is difficult to obtain better captions than *Dense Captioner* by only learning to use the union of subject and object boxes, despite having a larger number of captions.

We show example predictions of our relational captioning model in Fig. 4. Our model is able to generate rich and diverse captions for an image. We also show a comparison with the traditional frameworks, image captioner [33] and dense captioner [12]. While the dense captioner is able to generate diverse descriptions than an image captioner by virtue of various regions, our model can generate an even greater number of captions from the combination of the bounding boxes.

4.4. Comparison with Scene Graph

Motivated by scene graph, which is derived from the VRD task, we extend to a new type of a scene graph, which we call "caption graph." Figure 5 shows the caption graphs generated from our MTTSNet as well as the scene graphs from Neural Motifs [43]. For caption graph, we follow the same procedure as Neural Motifs but replace the relationship detection network into our MTTSNet. In both methods, we use ground truth bounding boxes to generate scene (and caption) graphs for fair comparison.

By virtue of being free form, our caption graph can have richer expression and information including attributes, whereas the traditional scene graph is limited to a closed set of the `subj-pred-obj` triplet. For example, in Fig. 5-(b,d), given the same object 'person,' our model is able to distinguish the fine-grained category (*i.e.* man vs boy and man vs woman). In addition, our model can provide more status information about the object (*e.g.* standing, black), by virtue of the attribute contained in our relational captioning data. Most importantly, the scene graph can contain unnatural relationships (*e.g.* tree-on-tree in Fig. 5-(c)), because prior relationship detection methods, *e.g.* [43], predict object classes individually. In contrast, by predicting the full sentence for every object pair, relational captioner can assign a more appropriate word for an object by considering the relations, *e.g.* "Green leaf on a tree."

Lastly, our model is able to assign different words for the same object by considering the context (the man vs baseball player in Fig. 5-(d)), whereas the scene graph generator can only assign one most likely class (man). Thus, our relational captioning framework enables more diverse interpretation of the objects compared to the traditional scene graph generation models.

	words/img	words/box
Image Cap. [33]	4.16	-
Scene Graph [43]	7.66	3.29
Dense Cap. [12]	18.41	4.59
Relational Cap. (**MTTSNet**)	**20.45**	**15.31**

Table 3: Diversity comparison between image captioning, scene graph generation, dense captioning, and relational captioning. We measure the number of different words per image (words/img) and the number of words per bounding box (words/box).

	R@1	R@5	R@10	Med
Full Image RNN[13]	0.10	0.30	0.43	13
Region RNN [8]	0.18	0.43	0.59	7
DenseCap [12]	0.27	0.53	0.67	5
RelCap (**MTTSNet**)	**0.29**	**0.60**	**0.73**	**4**

Table 4: Sentence based image retrieval performance compared to previous frameworks. We evaluate ranking using recall at k (R@K, higher is better) and the median rank of the target image (Med, lower is better).

4.5. Additional Analysis

Vocabulary Statistics. In addition, we measure the vocabulary statistics and compare them among the frameworks. The types of statistics measured are as follows: 1) an average number of unique words that have been used to describe an image, and 2) an average number of words to describe each box. More specifically, we count the number of unique words in all the predicted sentences and present the average number per image or box. Thus, the metric measures the amount of information we can obtain given an image or a fixed number of boxes. The comparison is depicted in Table 3. These statistics increase from *Image Cap.* to *Scene Graph* to *Dense Cap.* to *Relational Cap.* In conclusion, the proposed relational captioning is advantageous in diversity and amount of information, compared to both of the traditional object-centric scene understanding frameworks, scene graph generation and dense captioning.

Sentence-based Image and Region-pair Retrieval. Since our relational captioning framework produces richer image representations than other frameworks, it may have benefits on the sentence based image or region-pair retrieval, *which cannot be performed by scene graph generation or VRD models*. To evaluate on the retrieval task, we follow the same procedure as in Johnson *et al*. [12] with our relational captioning data. We randomly choose 1000 images from the test set, and from these chosen images, we collect 100 query sentences by sampling four random captions from 25 randomly chosen images. The task is to retrieve the correct image for each query by matching it with the generated captions.

We compute the ratio of the number of queries, of which the retrieved image ranked within top $k \in \{1, 5, 10\}$, and the total number of queries (denoted as *R@K*). We also report the median rank of the correctly retrieved images across all 1000 test images (The random chance performance is 0.001, 0.005, and 0.01 for R@1, R@5, and R@10 respectively). The retrieval results compared with several baselines are shown in Table 4. For baseline models *Full Image RNN*, *Region RNN*, and *DenseCap*, we display the performance measured from Johnson *et al*. [12]. To be compatible, we followed the same procedure of running through random test sets 3 times to report the average results. Our matching score is computed as follows. For every test image, we generate 100 region proposals from the RPN followed by NMS. In order to produce a matching score between a query and a region pair in the image, we compute the probability that the query text may occur from the region pair. Among all the scores for the region pairs from the image, we take the maximum matching score value as a representative score of the image. This score is used as the matching score between the query text and the image, and thus the images are sorted by rank based on these computed matching scores. As shown in Table 4, the proposed relational captioner outperforms all baseline frameworks. This is meaningful because region pair based method is more challenging than a single region based approaches.

5. Conclusion

We introduce relational captioning, a new notion which requires a model to localize regions of an image and describe each of the relational region pairs with a caption. To this end, we propose the MTTSNet, which facilitates POS aware relational captioning. In several sub-tasks, we empirically demonstrate the effectiveness of our framework over scene graph generation and the traditional captioning frameworks. As a way to represent imagery, the relational captioning can provide diverse, abundant, high-level and interpretable representations in caption form. In this regard, our work may open interesting applications, *e.g*., natural language based video summarization [3] may be benefited by our rich representation.

Acknowledgements. This work was supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No.2017-0-01780, The technology development for event recognition/relational reasoning and learning knowledge based system for video understanding)

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[2] Fuhai Chen, Rongrong Ji, Xiaoshuai Sun, Yongjian Wu, and Jinsong Su. Groupcap: Group-based image captioning with structured relevance and diversity constraints. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[3] Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. Contextually customized video summaries via natural language. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018. 8

[4] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. Towards diverse and natural image descriptions via a conditional gan. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017. 2

[5] Bo Dai and Dahua Lin. Contrastive learning for image captioning. In *Advances in Neural Information Processing Systems (NIPS)*, 2017. 2

[6] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *The workshop on statistical machine translation*, 2014. 5

[7] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[8] Ross Girshick. Fast R-CNN. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. 8

[9] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2

[10] Wenhao Jiang, Lin Ma, Yu-Gang Jiang, Wei Liu, and Tong Zhang. Recurrent fusion network for image captioning. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[11] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Inferring and executing programs for visual reasoning. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

[12] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2, 3, 4, 5, 6, 7, 8

[13] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2, 8

[14] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, Youngjin Yoon, and In So Kweon. Disjoint multi-task learning between heterogeneous human-centric tasks. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018. 1

[15] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*, 123(1):32–73, 2017. 1, 2, 5

[16] Michael F Land, Sophie M Furneaux, and Iain D Gilchrist. The organization of visually mediated actions in a subject without eye movements. *Neurocase*, 8(1):80–87, 2002. 1

[17] Yikang Li, Wanli Ouyang, and Xiaogang Wang. VIP-CNN: A visual phrase reasoning convolutional neural network for visual relationship detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[18] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and region captions. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*. Springer, 2014. 1

[20] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision (ECCV)*. Springer, 2016. 1, 2

[21] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[22] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1

[23] Aude Oliva and Antonio Torralba. The role of context in object recognition. *Trends in cognitive sciences*, 11(12):520–527, 2007. 1

[24] Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems (NIPS)*, 2011. 2

[25] Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. Weakly-supervised learning of visual relations. *IEEE International Conference on Computer Vision (ICCV)*, 2017. 3

[26] Bryan A Plummer, Arun Mallya, Christopher M Cervantes, Julia Hockenmaier, and Svetlana Lazebnik. Phrase localization and visual relationship detection with comprehensive linguistic cues. *IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. 2, 5

[28] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 6

[29] Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. Speaking the same language: Matching machine to human captions by adversarial training. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015. 3

[31] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 1

[32] Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond J Mooney, Trevor Darrell, and Kate Saenko. Captioning images with diverse objects. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[33] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2, 4, 6, 7, 8

[34] Liwei Wang, Alexander Schwing, and Svetlana Lazebnik. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In *Advances in Neural Information Processing Systems (NIPS)*, 2017. 2

[35] Sanghyun Woo, Dahun Kim, Donghyeon Cho, and In So Kweon. Linknet: Relational embedding for scene graph. In *Advances in Neural Information Processing Systems (NIPS)*, 2018. 2

[36] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[37] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML)*, 2015. 2

[38] Linjie Yang, Kevin Tang, Jianchao Yang, and Li-Jia Li. Dense captioning with joint inference and visual context. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 3, 4

[39] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[40] Guojun Yin, Lu Sheng, Bin Liu, Nenghai Yu, Xiaogang Wang, Jing Shao, and Chen Change Loy. Zoom-net: Mining deep feature interactions for visual relationship recognition. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[41] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[42] Ruichi Yu, Ang Li, Vlad I Morariu, and Larry S Davis. Visual relationship detection with internal and external linguistic knowledge distillation. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[43] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 5, 6, 7, 8