symmetry

MDPI

Article

Image Caption Generation Using Multi-Level Semantic Context Information

Peng Tian, Hongwei Mo * and Laihao Jiang

College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin 150001, China; tianpengheu@hrbeu.edu.cn (P.T.); lhjiang@hrbeu.edu.cn (L.J.)
* Correspondence: honwei2004@126.com

Abstract: Object detection, visual relationship detection, and image captioning, which are the three main visual tasks in scene understanding, are highly correlated and correspond to different semantic levels of scene image. However, the existing captioning methods convert the extracted image features into description text, and the obtained results are not satisfactory. In this work, we propose a Multi-level Semantic Context Information (MSCI) network with an overall symmetrical structure to leverage the mutual connections across the three different semantic layers and extract the context information between them, to solve jointly the three vision tasks for achieving the accurate and comprehensive description of the scene image. The model uses a feature refining structure to mutual connections and iteratively updates the different semantic features of the image. Then a context information extraction network is used to extract the context information between the three different semantic layers, and an attention mechanism is introduced to improve the accuracy of image captioning while using the context information between the different semantic layers to improve the accuracy of object detection and relationship detection. Experiments on the VRD and COCO datasets demonstrate that our proposed model can leverage the context information between semantic layers to improve the accuracy of those visual tasks generation.

Keywords: scene understanding; object detection; visual relationship; image captioning; semantic level; context information; attention mechanism

Citation: Tian, P.; Mo, H.; Jiang, L. Image Caption Generation Using Multi-Level Semantic Context Information. *Symmetry* **2021**, *13*, 1184. https://doi.org/10.3390/sym13071184

Academic Editor: Dumitru Baleanu

Received: 27 May 2021
Accepted: 28 June 2021
Published: 30 June 2021

1. Introduction

Image captioning is the research hotspot of computer vision technology, and it is one of the main tasks to realize the scene understanding. Image captioning is based on detecting and recognizing objects, reasoning the relationship of the detected objects, and using natural language to describe the semantic content of the scene image [1,2]. Image and description text are two different representation manners, they are symmetric and unified in the semantic content of the same visual scene. The traditional image-captioning methods convert the extracted image features into the description text, which can only achieve a simple description of the scene image. In recent years, with the rapid development of computer vision and deep learning technology, the surface features of image captioning is receiving less attention from the research community, and instead the focus is on accurate and comprehensive deep image-captioning research [3,4].

Image captioning not only needs to detect objects and reason the relationships between them, but also needs to consider the context information between the target objects, and object and the image scene environment to achieve the accurate and comprehensive description of the main semantic content of the image, and show the differences in content displayed by different scene images. Therefore, image captioning has high complexity. As shown in Figure 1, the two images with different semantic content can be simply described as "a person is laying," when we do not consider the context information between the person and other objects, as well as between the person and the image scene environment. By contrast, when we take the context information between the person and bed, and person

and room in this image on the left of Figure 1 into account, we are more likely to get a caption as "a person is sleeping on a bed in a room". Obviously, the context information has an important influence on the description word generation. Therefore, ignoring the context information inevitably places constraints on the accuracy and comprehensiveness of the semantic content of the image described. The key to image captioning is to detect and recognize objects and their relationships accurately, and to obtain the context information between objects and relationships, and objects and image scene. This information not only helps for achieve a deeper image understanding, but also provide values for high-level vision understanding and reasoning tasks [5,6]. To describe the scene image more accurately and comprehensively, we need to jointly solve the three visual tasks of object detection, relationship detection, and image captioning, to leverage the context information between the different semantic layers to promote the generation of visual tasks.

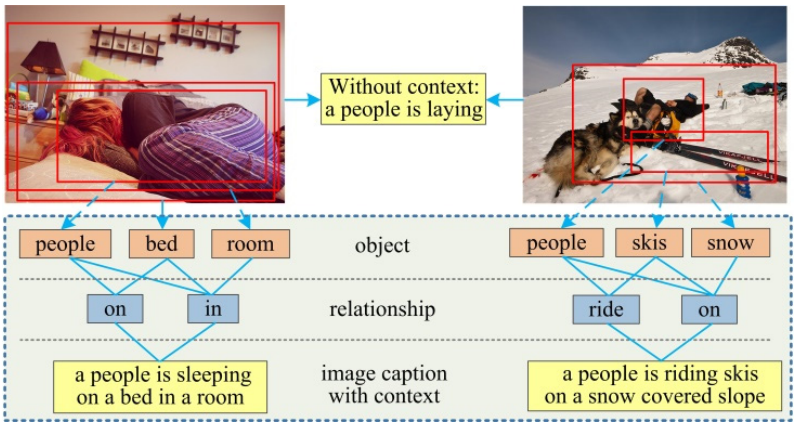

Figure 1. Context information affects the generation of description words. For images with different semantic content, ignoring the context information related to the target object, the image captioning would generate the same result (**top**). We propose an image-captioning model based on multi-level semantic context information, which leverages the context information between different semantic layers in the image to generate accurate and comprehensive description sentences (**bottom**).

Compared with the earlier image-captioning methods, the modern captioning methods based on the encoder-decoder neural network framework first encode the input image as the vector feature, and then decode the vector feature into a descriptive sentence, and the description performance is improved [7–10]. A series of approaches [2,4,7] used the neural networks to obtain the image features to generate description sentences. Other methods [11–13] attempted to leverage attention mechanisms to focus on the features of different regions of the scene image to improve description performance. Some other works [14,15] introduced the scene graph into image captioning to further improve the description performance. However, these methods generate captions by extracting the image global features, or focusing on the context information between the target objects in the image to detect the object and reason the relationship, and then according to the corresponding grammatical structure generate the caption sentences by using the fixed phrases or conjunctions to combine object and relationship words. In addition, these works do not take the context information between objects and relationships, and objects and image scene into account, which may lead to the generated description sentence that cannot accurately and comprehensively describe the main semantic content of the image. Different from these captioning methods, our method jointly solves the visual tasks corresponding to different semantic layers in the scene image, and leverages the context information between the different semantic layers to simultaneously promote the generation of object detection, relationship detection, and image-captioning tasks.

In this work, we propose a Multi-level Semantic Context Information (MSCI) network, to solve jointly the object detection, reason the relationship between the detected objects, and image-captioning vision tasks, to leverage the mutual connections and iteratively

update the features of the three different semantic layers and extract the context information between the different semantic layers to simultaneously promote the generation of the three visual tasks for achieving the accurate and comprehensive description of the scene image. The experiments on the VRD and COCO datasets demonstrate that our proposed model can simultaneously improve the accuracy of object detection, relationship detection, and image captioning, and the superiority of caption performance over the other existing captioning models.

The contributions of our work are summarized as follows:

- We propose a multi-level semantic context information network model to simultaneously solve the object detection, visual relationship detection, and image-captioning tasks.
- We use the Gated Graph Neural Network (GGNN) [16] to pass and refine the features of different semantic layers, and then build a context information extraction network to respectively obtain the relationship context between objects and the scene context between object and image scene, the model can leverage the context information between the different semantic layers to simultaneously improve the accuracy of the visual tasks generation.
- We use the extracted context information as the input of the attention mechanism, and integrate its output into the decoder with the caption region features. In this way, the model can obtain more image features and while focus on the context information between objects, relationships, and image scene to improve the accuracy and comprehensiveness of the generated caption.

2. Related Work

2.1. Object Detection

Object detection is one of the foundation tasks to achieve visual understanding and reasoning. Girshick et al. [17] first used convolutional neural networks (CNN) for object detection, and then researchers proposed a variety of CNN-based detection algorithms to improve the detection effect, such as Faster R-CNN [18], YOLOV4 [19], and SSD [20]. The detection speed of YOLO network is fast, but the accuracy of object position regression is insufficient, SSD makes the network structure complex by sharing multiple convolutional layers. Faster R-CNN uses region proposal network (RPN) to generate candidate region proposals, which has high accuracy of object position regression and detection efficiency. Therefore, we use Faster R-CNN for object detection.

2.2. Visual Relationship Detection

Visual relationship detection is one of the core tasks for image captioning. It represents the state of the relationship between objects, which can be represented as a (subject-predicate-object) phrase with structure symmetric. In early days, researchers used specific types of visual phrases to detect relationships. Recently, most of the existing pipelines have leveraged the interaction information between objects to improve relationship detection [21–23]. Lu et al. [24] used the visual features of object and the semantic similarity between descriptive words for relationship recognition. Dai et al. [25] used the location information of objects to identify the location relationship. Chen et al. [26] used the statistical information between objects and their relationships to boost relationship detection. However, these works are only reasoning the visual relationship by extracting the correlation features, and not establishing connection between visual relationships and other vision tasks such as image captioning. Different from these works, we will view the objects, visual relationships, and image captioning as three different semantic layers of scene understanding, and try to reason the visual relationships through mutual connections and iteratively update the features of the three different semantic layers and extract the context information from these semantic layers.

2.3. Image Captioning

Image captioning uses natural language to represent the objects, their attributes, and relationships in an image [6]. Compared with the early image-captioning methods, template-based and retrieval-based [27–30], the modern image-captioning methods based on encoder-decoder structure have made great progress [31,32]. Inspired by machine translation, researchers have proposed many image-captioning models based on attention mechanism to improve the description performance [33,34]. Other approaches [35,36] also attempted to leverage the scene graph to capture the relationship information between objects to boost the performance of image captioning. However, these algorithms are mainly based on the visual features of the target objects to guide the generation of description sentences, and it has two major drawbacks. First, they rarely consider the spatial information with the target objects, which may lead to the inaccurate description of the relationship between the target objects. Second, they do not consider the image scene information, which may lead to inaccurate description of the semantic content of the entire image. Our work is to improve the description performance of the captioning model by extracting the relationship context information and scene context information correlated to the target objects in the image and combining the attention mechanism to obtain more useful description features. The whole structure of our proposed method is symmetrical, and more feature information can be extracted from the input image to boost image captioning.

2.4. Context Information

Context information represents the interaction between objects or between objects and the surrounding environment. The rich context information in the scene image provides values for relationship detection and image-captioning tasks [37]. Early context research is mainly applied to object detection [38]. Recently, with the rapid development of deep learning, scholars have tried to use neural network-based methods to extract context information from images and apply it to high-level visual tasks [39–41]. Zhuang et al. [42] utilized a context-interaction classification structure to detect the vision relationship between objects. Zellers et al. [43] used the object context for scene graph generation. These methods only used the extracted context information for a single vision task, and not consider extracting the context information between different semantic layers in image and used it to jointly solve multiple vision tasks. Different from these works, we propose a context information extraction network with structure symmetry that simultaneously models the relationship context between objects and the scene context between object and image scene, and leverages the extracted context information to simultaneously promote the generation of object detection, visual relationship detection, and image-captioning tasks.

3. Method

3.1. Model Overview

We propose a Multi-level Semantic Context Information (MSCI) model, which through a feature update structure jointly refines the features of the object, relationship, and caption semantic layers, and leverages the context information between the different semantic layers to improve the accuracy and comprehensiveness of the generated caption. The MSCI model mainly consists of four parts: feature extraction, feature update, context information extraction, and caption generation. The whole framework of our model (MSCI) is shown in Figure 2, and it is a symmetric structure.

The entire process of the model is summarized as follows: (1) Region proposal and feature extraction. To generate region proposals of objects, relationship phrases and image captions, and extract the features of each region for different semantic tasks. (2) Feature updating. To mutual connect the features of object, relationship, and caption semantic layers based on their spatial relationships, and then jointly iteratively update these features by the GGNN [16]. (3) Context information extraction. A context information extraction network is used to extract relationship context information and scene context information

from the updated semantic features, and then take the context information as the input of the attention mechanism. (4) Image caption generation. The output of the attention mechanism and the caption features are merged into the decoding structure to generate description sentences. The core of our work is to use the GGNN for simultaneously passing and updating the features of different semantic layers, and introduce a context extraction network to extract the context information between the different semantic layers to simultaneously improve the accuracy of object detection and relationship detection, and the accuracy and comprehensiveness of the generated image caption.

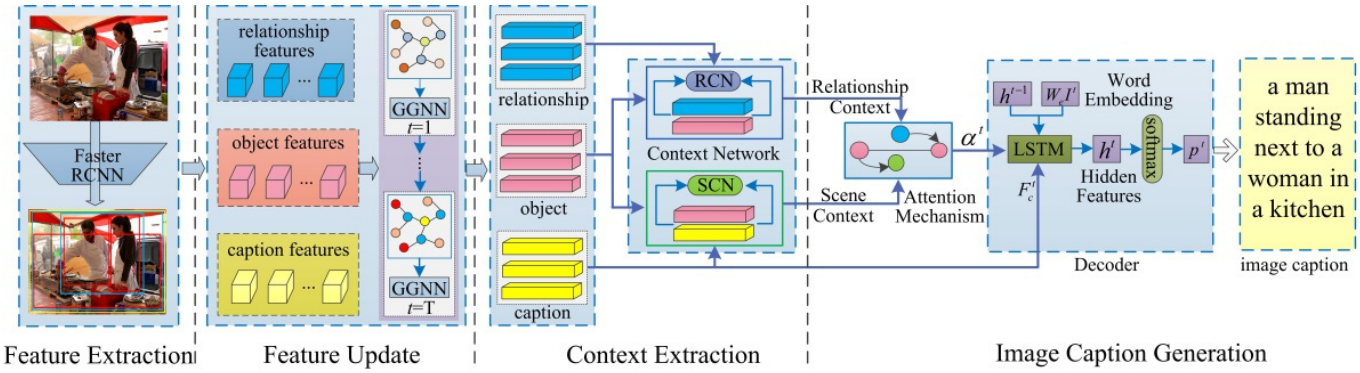

Figure 2. The whole framework of MSCI model. We first adopt the Faster R-CNN to detect a set of object regions and caption regions, and the relationship regions are generated by grouping and connecting object regions into object pairs, ROI-pooling is used for obtaining the features of object, relationship, and caption region proposals, and then these feature messages are fed into the GGNN for pass and update. After message passing, the features of object, relationship, and caption are fed into the relationship context network (RCN) and scene context network (SCN) to extract the relationship context information and scene context information, respectively. Then the updated object features and the extracted relationship context information are used for object detection and relationship detection. Finally, the model takes the extracted relationship context information and scene context information as the input of an attention mechanism, and its output information and the caption features are merged into the decoder to generate caption sentence. Different colors are assigned to the features of object (red), relationship (blue), and caption (yellow).

3.2. Region Proposal and Feature Extraction

We first utilize Faster R-CNN [18] to extract the visual features of the input image and use them as the input of a Region Proposal Network (RPN) to generate the object region proposal set $B = \{b_1, b_2, \ldots, b_n\}$. For each object region, the model generates a bounding box b_i denoting the position of the object. On the other hand, we use another RPN trained with the ground truth region bounding boxes to generate the caption region proposals. We apply non-maximum suppression (NMS) [17] to reduce the ROI (region of interest) number of object and caption proposals, and filter out the overlap objects that the IOU of their bounding boxes is greater than 0.8. Finally, the object proposals are combined to form the relationship region proposals, and the ROI pooling layer is used to extract the object, relationship, and caption features corresponding to each proposal region, respectively.

3.3. Feature Updating

We use the model $I = \{B, V, S\}$ to represent the scene image, where B represents the region proposal set, V represents the node set, S represents the scene of the image. $b^o, b^r, b^c \in B$ denote the object regions, relationship regions and caption regions, and $v^o, v^r, v^c \in V$ denote the nodes of object, relationship, and caption, respectively.

The core of each node $v \in V$ in the model is the ability to encode features from other neighboring nodes and propagate its messages to neighbor nodes to achieve feature update. Therefore, we use a memory and propagate a structure that can memorize the node messages and simultaneously integrate the information from other nodes. Gated Graph Neural Network (GGNN) [16] is an end-to-end trainable network architecture that

can not only memorize the node features, but also integrate the received information from other neighboring nodes, and update the node representation in a recurrent fashion. Some works have successfully used the GGNN for visual tasks including semantic segmentation and image recognition [44,45]. We will view the proposal regions in the image as nodes and connect them based on their spatial relationships, and then leverage the GGNN to propagate and update the features of object, relationship, and caption nodes to explore the interaction and context information between objects, and object and image scene. For each node corresponding to the region that is detected in the image, we use the features extracted from the corresponding region to initialize the input feature of the node, and otherwise, it is initialized by a d-dimension zero vector. Supposing $f_i^o, f_i^r, f_i^c \in R^D$ denotes the feature vector of object, relationship, and caption region, respectively. The message passing and updating procedure of object regions, relationship regions and caption regions can be written as follows.

At each time step t, each node first aggregates messages from its neighbor nodes, which can be expressed as:

$$m_v^t = \begin{cases} \sum\limits_{j \in N_i^o} (f_j^o,\ f_j^r,\ h_j^{t-1}) \text{ if } v \text{ is the object node } o_i \\ \sum\limits_{j \in N_i^o} (f_i^o,\ f_j^o,\ h_j^{t-1}) \text{ if } v \text{ is the relationship node } r_i \\ \left[\sum\limits_{i=1}^{H} (f_i^o,\ h_j^{t-1}), \sum\limits_{i=1}^{L} (f_i^r,\ h_i^{t-1}) \right] \text{ if } v \text{ is the relationship node } c_i \end{cases} \tag{1}$$

where $f_j^o,\ f_j^r$ represent the feature messages of object node and relationship node respectively. N_i^v stands for the neighbors of node i. Each relationship node will be connected to two object nodes, and each caption node will be connected to multiple object nodes and relationship nodes. Then, the model incorporates information m_v^t from the other neighbor nodes and its previous hidden state as input to update its hidden state at the current moment by a gated mechanism similar to the gated recurrent unit [16].

We first compute the reset gate r

$$r_v^t = \sigma\left(W^r m_v^t + U^r h_v^{t-1}\right) \tag{2}$$

where σ represents the sigmoid activation function, W^r and U^r respectively represent the weight matrix, m_v^t represents the information received by node v at time step t, and h_v^{t-1} represents the hidden state of node v at the previous moment. Similarly, the update gate z is computed by

$$z_v^t = \sigma\left(W^z m_v^t + U^z h_v^{t-1}\right) \tag{3}$$

The activation function of the unit h_v^t can be expressed as:

$$h_v^t = \left(1 - z_v^t\right) \odot h_v^{t-1} + z_v^t \odot \widetilde{h}_v^t \tag{4}$$

where:

$$\widetilde{h}_v^t = \tanh\left(W m_v^t + U\left(\ r_v^t \odot h_v^{t-1}\right)\right) \tag{5}$$

where $\odot$ denotes the element-wise multiplication, tanh is the tangent active function, W and U are the learned weight matrixes, respectively. In this way, each node can aggregate messages from its neighboring nodes, and propagate its message to other neighboring nodes. After iterations T_v times, we can obtain the final hidden state $\left\{h_{i1}^{T_v}, h_{i2}^{T_v}, \ldots, h_{iL}^{T_v}\right\}$, $L \in (N, M, K)$ for each node. We adopt a fully connected network layer φ to compute the output feature for each node, and the network takes the initial hidden state h_{il}^0 and final hidden state $h_{iL}^{T_v}$ of

the node as input. For each proposed region, we aggregate all correlated output features of the node and express it as the feature F_v of the region.

$$F_v = \begin{cases} \sum_{i=1}^{N} \varphi_o(h_o^0, h_i^{To}) \text{ if } v \text{ is the object node } o \\ \sum_{i=1}^{M} \varphi_r(h_r^0, h_i^{Tr}) \text{ if } v \text{ is the relationship node } r \\ \sum_{i=1}^{K} \varphi_c(h_c^0, h_i^{Tc}) \text{ if } v \text{ is the relationship node } c \end{cases} \quad (6)$$

For the obtained features F_o of each object region, we send it into an object classification fully connected network layer to predict the class label of the object. For the obtained features F_r of each relationship region, we use a relationship classification fully connected network layer that takes it and the extracted relationship context in the subsequent network as input to predict the class label of the relationship. The updated features F_c of caption region and the extracted context information in subsequent network are fed into the captioning model to generate the description sentence. With this structure of feature passing and updating, each node can aggregate messages from its neighboring nodes and propagate its messages to the neighboring nodes to achieve feature update. The updated features will be used to extract context information and generate captions.

3.4. Context Information Extraction

The core of our work is to improve the accuracy and comprehensiveness of the model-generation caption by exploring rich context information between different semantic layers in the scene image. Different visual tasks in scene understanding correspond to different semantic layers of the scene image; the image features of each semantic layer only contains the corresponding visual features, and not the context information between different semantic layers and the spatial structure information of the image. Therefore, in order to accurately and comprehensively describe the semantic content of an image, we need to detect and extract the visual features of the candidate regions and the context information correlated to the visual relationship and image captioning. In order to incorporate the spatial structure information of the scene image into the context information, we introduce the relative spatial position information of object–object and object–scene in the captioning model to improve the accuracy of the generated caption. Figure 3 illustrates the symmetric framework of the context information extraction network, we use the network to extract the relationship context information c_r between objects for relationship detection and the scene context information c_s between object and image scene for image captioning.

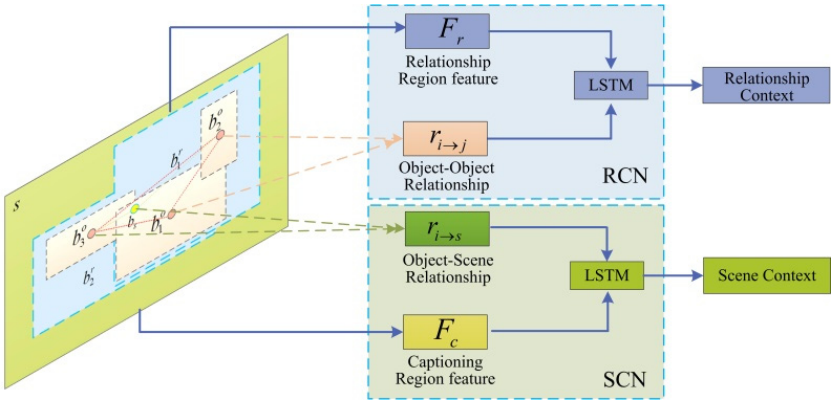

Figure 3. Context information extraction network.

3.4.1. Relationship Context

We use the feature F_r of the relationship region and the relationship information between the objects as the input of the relationship context network (RCN) to extract the relationship context information c_r. The relationship between objects is inferred through their spatial position information and can enrich the context information. The object–object relationship information $r_{i\to j}$ is determined by the relative object position and the visual relationship features; the visual relationship vector is formed by concatenating the object features F_{o_i} and F_{o_j}.

In the context extraction network, we use a Long Short-Term Memory (LSTM) model [46] to extract context information. The relationship feature F_r is taken as the initial state of the LSTM, the relationship information $r_{i\to j}$ is used as the input of the LSTM, and the output is relationship context information c_r.

$$c_r = LSTM\left(F_r, r_{i\to j}\right) \tag{7}$$

$$r_{i\to j} = relu\left(W_b R_{i\to j}\left(b_i, b_j\right)\right) * \tanh\left(W_r\left[F_{o_i}, F_{o_j}\right]\right) \tag{8}$$

$$R_{i\to j}\left(b_i, b_j\right) = \left[\frac{x_i - x_j}{w_j}, \frac{y_i - y_j}{h_j}, log\frac{w_i}{w_j}, log\frac{h_i}{h_j}, \frac{a_i \cap a_j}{a_i \cup a_j}\right] \tag{9}$$

where W_b and W_v are weight matrixes which are learned, $R_{i\to j}$ represents the relative spatial position information of object pairs, b_i, b_j represent the object proposal that form the relationship proposal, respectively. (x_i, y_i) denotes the center of object proposal bounding box b_i, w_i, and h_i denote the width and height of b_i and a_i represents the area of b_i. For the obtained relationship context information c_r, we take the features F_r of the corresponding relationship region as input of a classification fully connected network layer to predict the class label of the relationship.

3.4.2. Scene Context

We use the caption region feature F_c and the relationship information of the object relative to image scene as the input of the scene context network (SCN) to extract the scene context information c_s. The relationship between object and scene is inferred through their spatial position information and enrich the context information. The relationship information $r_{i\to s}$ between object and image scene is determined by the relative position and visual relationship features. The visual relationship vector is formed by concatenating the object feature F_{o_i} and the caption region feature F_c.

The caption region feature F_c is taken as the initial state of the another LSTM, the relationship information $r_{i\to s}$ of the object relative to the image scene is used as the input of the LSTM, and the output is the scene context information c_s. Then the relationship context and scene context are used as the input of the attention mechanism.

$$c_s = LSTM(F_c, r_{i\to s}) \tag{10}$$

$$r_{i\to s} = relu(W_b R_{i\to s}(b_i, b_s)) * \tanh(W_s[F_{o_i}, F_c]) \tag{11}$$

$$R_{i\to s}(b_i, b_s) = \left[w_i,\ h_i,\ w_s, h_s,\ \frac{x_i - x_s}{w_s}, \frac{y_i - y_s}{h_s}, log\frac{w_i}{w_s}, log\frac{h_i}{h_s}\right] \tag{12}$$

where W_s is weight matrixes which is learned, $R_{i\to s}$ denotes the spatial position information of the object relative to the image scene, b_s represents the caption proposal, (x_s, y_s) is the center of the caption proposal bounding box b_s, w_s, h_s are the width and height of b_s.

3.5. Image Captioning

Image captioning uses natural language to generate description sentences corresponding to the semantic content of the image. We use a spatial attention mechanism that takes the relationship context information and scene context information as input, and then the

attention mechanism dynamically selects image features related to word generation at the current moment from the feature vector of caption region according to the hidden state of the LSTM network to guide word generation. The model selects the relationship context information c_r, scene context information c_s, and the previous hidden state h^{t-1} of the LSTM network in the decoder to determine the weight α_i^t of the relationship context information and scene context information corresponding to the target object at the current moment.

$$e_i^t = f_{att}\left(c_r^t, c_s^t, h^{t-1}\right) \tag{13}$$

$$\alpha_i^t = \frac{exp(e_i^t)}{\sum_{k=1}^{N} exp(e_k^t)} \tag{14}$$

where f_{att} is an attention mechanism model, and its output results are normalized to obtain the weight distribution of context information at time step t, these weight denotes the captioning model pay attention to the context information.

The input information I^t of the captioning model at the current time includes the caption feature F_c^t, the relationship context information c_r^t, the scene context information c_s^t, and their corresponding attention weight information α_i^t.

$$I^t = \left\{ \sum_{i=1}^{N} \alpha_i^t (c_r^t, c_s^t), F_c^t \right\} \tag{15}$$

The model uses the output y^{t-1} and hidden state h^{t-1} at the previous moment and the input information I^t at the current moment to calculate the hidden state h^t of the LSTM at the current moment, then according to the hidden state h^t, the output y^{t-1} at the previous moment and the input information I^t at the current moment, the probability distribution of the output word at the current moment is obtained by Softmax:

$$h^t = LSTM\left(y^{t-1}, h^{t-1}, I^t\right) \tag{16}$$

$$p(y^t \left| y^{t-1}\right.) \ = \ softmax\left(y^{t-1}, h^t, I^t\right) \tag{17}$$

We first use the cross-entropy loss to train the proposed model. Given the ground-truth captions m, and θ represents the parameter in the model, the cross-entropy loss function $L(\theta)$ can be written as:

$$L_{XE}(\theta) = -\sum_{t=1}^{T} log(p_\theta(m_t|m_{t-1})) \tag{18}$$

where $p_\theta(m_t|m_{t-1})$ represents the output probability of the word m_t. Subsequently, we introduce the reinforcement learning (RL) method [13,33] to optimize the model to maximize the expected reward:

$$L_{RL}(\theta) = -\sum_{i=1}^{N} E_{s^i \sim p_{\theta_i}}\left[r\left(s^i\right)\right] \tag{19}$$

where r is the CIDEr reward function, s^i represents the word sequence sampled by the model, and p_{θ_i} represents the probability distribution of the words. Finally, the gradient of each training sample can be approximated:

$$\nabla_\theta \, L_{RL}(\theta) \ \approx -\sum_{i=1}^{N} r\left(s^i\right) \cdot \nabla_{\theta_i} log p_{\theta_i}\left(s^i\right) \tag{20}$$

The token <S> indicates the model start to generate the description sentence, and selects the word with the highest probability distribution as the output word at the current

moment. The words are generated in a recurrent fashion until the stop token <E> or the maximum length of the word sequence is reached.

4. Experiments

4.1. Datasets and Evaluation Metrics

4.1.1. Datasets

To verify the effectiveness of our proposed model, we evaluate the performance of the proposed model on the COCO [47] and Flickr30k [48] benchmark datasets. COCO is a standard dataset for the current research on image captioning. In this dataset, there are 164,062 images that have the captions, and each image has five different reference captions annotated by human. Karpathy [2] split the dataset, and used 113,287 images and corresponding reference captions as the training set, then the 5000 images and corresponding reference captions are selected as the testing set and validation set, respectively. Flickr30k dataset consists of 31,783 images, and each image has five different reference captions annotated by human. It contains 7000 words and 158,915 captions. Following the previous work [2], we use 1000 images for validation, 1000 images for testing, and the rest images for training.

Before training the model, all the words in the reference captions are changed to lower case, the punctuation replaced with spaces, and the words that appear less than five times were deleted to avoid rare words that are not conducive to the generation of description text. In addition, we evaluate the relationship detection performance of our model on the Visual Relationship Detection (VRD) standard dataset [24]. This dataset contains 5000 images with 100 object categories and 70 relationship predicates. In our experiments, we followed [24], and used 4000 images for training and 1000 images for testing.

4.1.2. Evaluation Metrics

We use BLEU-N(B@N) [49], METEOR(M) [50], ROUGE(R) [51], CIDEr(C) [52], and SPICE(S) [53] the five metrics to evaluate the performance of our proposed model. We use the evaluation metrics Recall@50 and Recall@100 [24] for relationship detection.

4.2. Implementation Details

We adopt the Faster R-CNN detector with VGG-16 pre-trained on ImageNet to generate the candidate region proposals. We first train RPN and set the NMS threshold to 0.7 for region proposals, and use the SGD algorithm training detector with a batch size of 6 and momentum of 0.9, and the model is trained with initializing the learning rate as 0.008. After that, we first freeze the parameters of convolution layers and then train the models of relationship detection and image captioning. The iteration time of the GGNN model is set as 5, the dimension of the hidden state is set as 2048, and the dimension of the output feature is set as 512. We use the cross-entropy loss to train the relationship detection model on the VRD dataset and the learning rate is set to 0.0005. For the captioning model, we first train the model with cross-entropy loss, and then train it with the reinforcement learning optimizing CIDEr on the COCO dataset. We used the training optimizer similar to Adam's [54], the learning rate is set to 0.0001, and the batch is set as 32. The dimensions of the image feature and object feature are set as 2048, the dimension of the hidden layer in the LSTM network and word vector are set as 512, and the maximum length of the generated captioning text is set as 16.

4.3. Performance Comparison and Analysis

4.3.1. Compared Methods

Our proposed MSCI model and the following existing state-of-the-art models use ResNet-101 as the backbone of the encoder, and are tested on the same COCO dataset for fair comparison. (1) The DA [13] model devises a residual attention network to obtain more feature information to facilitate image description. (2) Trans-KG [8] model leverages the knowledge graph to improve the image captioning. (3) ORRA [36] model introduces

a hierarchical attention module to learn discriminative features of words generation for image captioning. (4) Sub-GC [15] model explores the representation of scene graph to generate high-quality captions. (5) Up-Down [33] model uses a bottom-up attention mechanism to obtain the attention information to improve the performance of the model-generation caption. (6) HAN [34] model uses the hierarchy features to predict different words. (7) STMA [32] model uses an attention model to learn the spatio-temporal relationships for image captioning. (8) SGAE [14] model introduces the language induction bias into the captioning framework to obtain the human-like descriptions. (9) MCTG [31] model introduces a control signal to control the sentence attributes for enhancing the diversity of generated captions. (10) VRAtt-Soft [9] model integrates the visual relationship attention and visual region features to boost image captioning.

4.3.2. Performance Analysis

We compare the performance of our proposed MSCI model with the state-of-the-art captioning models on the COCO dataset, and the results are summarized in Table 1.

Table 1. The performance of the MSCI model and other captioning models on COCO Karpathy test split.

Model	**Cross-Entropy Loss**						**CIDEr Score Optimization**					
Metric	**B@1**	**B@4**	**M**	**R**	**C**	**S**	**B@1**	**B@4**	**M**	**R**	**C**	**S**
DA [13]	0.758	0.357	0.274	0.562	1.119	0.205	0.799	0.375	0.285	0.582	1.256	0.223
Trans-KG [8]	0.764	0.344	0.277	0.565	1.128	0.209	-	-	-	-	-	-
ORRA [36]	0.767	0.338	0.262	0.549	1.103	0.198	0.792	0.363	0.276	0.568	1.202	0.214
Sub-GC [15]	0.768	0.362	0.277	0.566	1.153	0.207	-	-	-	-	-	-
Up-Down [33]	0.772	0.362	0.270	0.564	1.135	0.203	0.798	0.363	0.277	0.569	1.201	0.214
HAN [34]	0.772	0.362	0.275	0.566	1.148	0.206	0.809	0.376	0.278	0.581	1.217	0.215
STMA [32]	0.774	0.365	0.274	0.568	1.144	0.205	0.802	0.377	0.282	0.581	1.259	0.217
SGAE [14]	0.776	0.369	0.277	0.572	1.167	0.209	0.808	0.384	0.284	0.586	1.278	0.221
MCTG [31]	0.783	0.375	0.281	0.574	1.203	0.215	0.787	0.371	0.277	0.561	1.264	0.215
VRAtt-Soft [9]	0.792	0.369	0.283	0.609	1.143	0.208	0.804	0.375	0.285	0.616	1.221	0.221
MSCI(Ours)	0.793	0.378	0.292	0.586	1.186	0.216	0.812	0.393	0.295	0.593	1.275	0.225

We report the performance of our MSCI model and the state-of-the-art compared models on the test portion of the Karpathy splits in Table 1. The experimental results with the cross-entropy loss function to train the model are shown on the left, we can see that the scores of MSCI model outperform the compared image-captioning models on most evaluation metrics. This shows that MSCI model can effectively improve the performance of model-generation captions by leveraging the context information between the different semantic layers, and the superiority of MSCI model-generation captions over the existing image-captioning models. Compared with the VRAtt-Soft model, the scores of the MSCI model on BLEU-1, BLEU-4, METEOR, CIDEr, and SPICE metrics increased by 0.001 ~ 0.043, and the score on the ROUGE metric is slightly lower than the VRAtt-Soft model by 0.023. Compared with the MCTG model, the scores of the MSCI model on the other evaluation metrics increased 0.001 ~ 0.012, and the score on the CIDEr metric is slightly lower than the MCTG model by 0.017. The above discussion and comparison results confirm that the MSCI model refines the features of object, relationship, and caption, and extract the relationship context information and scene context information in the scene image can improve the performance of the model-generation caption.

For fair comparison the performance of our MSCI model, we also report the performance of MSCI model optimized with CIDEr score in the right of Table 1. Focusing on the comparison results, we can see that the scores of the MSCI model optimized by the CIDEr score on all metrics increased by 0.003 ~ 0.089, the score on the CIDEr metric improved from 1.186 to 1.275, and the scores of MSCI model outperform the compared image-captioning models optimized by the CIDEr score on most evaluation metrics. The comparison results show that MSCI model optimized with the CIDEr score can significantly promote the

score on each evaluation metric, and the performance of MSCI model outperforms other captioning models.

We also validate our model on Flicker30k dataset with comparison to the state-of-the-art methods. The experimental results with the cross-entropy loss function to train the model are shown on Table 2.

Table 2. The performance of the MSCI model and other captioning models on Flicker30k dataset.

Model	B@1	B@2	B@3	B@4	M	R	C	S
PoS [55]	0.638	0.446	0.307	0.211	-	-	-	-
Trans-KG [8]	0.676	-	-	0.260	0.219	-	0.575	0.163
Adaptive-ATT [12]	0.677	0.494	0.354	0.251	0.204	-	0.531	-
r-GRU [56]	0.694	0.458	0.336	0.231	0.298	0.442	0.479	-
Sub-GC [15]	0.707	-	-	0.285	0.223	-	0.619	0.164
MSCI(Ours)	0.705	0.503	0.374	0.294	0.238	0.457	0.613	0.167

From the experiment results in Table 2, we can see that our MSCI model performs better than the compared image-captioning models on multiple evaluation metrics. Specifically, compared with the Sub-GC model, the scores of the MSCI model on the other evaluation metrics increased 0.003 ~ 0.015, and the score on the BLUE-1 and CIDEr metrics are slightly lower than the Sub-GC model by 0.002 and 0.006, respectively. Compared with the r-GRU model, the score of the MSCI model on the METEOR metric is slightly lower than the r-GRU model by 0.06, and the scores on the other evaluation metrics increased 0.011 ~ 0.134. The above discussion and comparison results show that our proposed MSCI model can improve the performance of the model-generation caption by refining image features and extracting contextual information between different semantic layers in the image.

In addition, we also report the performance of our MSCI model and other state-of-the-art models trained with CIDEr score optimization on the official COCO evaluation server in Table 3. From the comparison results, we can see that MSCI model outperforms other models on most evaluation metrics.

Table 3. The results of MSCI model and other captioning models on the online COCO test sever.

Model	B@1		B@2		B@4		METER		ROUGE		CIDEr	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
DA [13]	0.794	0.944	0.635	0.880	0.368	0.674	0.282	0.370	0.577	0.722	1.205	1.220
ORRA [36]	0.792	0.944	0.626	0.872	0.354	0.658	0.273	0.361	0.562	0.712	1.151	1.173
Up-Down [33]	0.802	0.952	0.641	0.888	0.369	0.685	0.276	0.367	0.571	0.724	1.179	1.205
STMA [32]	0.803	0.948	0.646	0.888	0.377	0.686	0.283	0.371	0.581	0.728	1.231	1.255
HAN [34]	0.804	0.945	0.638	0.877	0.365	0.668	0.274	0.361	0.573	0.719	1.152	1.182
SGAE [14]	0.806	0.950	0.650	0.889	0.378	0.687	0.281	0.370	0.582	0.731	1.227	1.255
MSCI (Ours)	0.806	0.949	0.652	0.897	0.389	0.706	0.293	0.382	0.589	0.742	1.221	1.248

4.3.3. Qualitative Examples

We compare the description effect of the MSCI model with the Up-Down model and SGAE model on the same COCO and Flicker30k datasets. The caption examples generated by the models are shown in Figure 4.

From the caption examples in Figure 4, we can see that the captions generated by the MSCI model are closest to the annotated reference captions, the generated captions are accurate and comprehensive, and the superiority of MSCI model's description effect over the Up-Down model and SGAE model. MSCI model generates accurate and comprehensive captions by iteratively refining image features and extracting the contextual information between different semantic layers. The generated captions not only contain more descriptive information about the objects and relationships between the main objects, but also describe the semantic information between the target object and the image scene. For example, in

Figure 4a, the MSCI model describes the semantic information of the entire scene image: a large elephant is standing in a zoo enclosure; in Figure 4c, the phrase "on the back of" is used to describe the spatial relationship between man and horse, and the word "brown" is used to describe the color attribute of horse; in Figure 4d, the phrase "standing on" is used to describe the relationship between target object "girl" and the image scene "sidewalk". In Figure 4e, when considering the scene context information, the captioning generated by the MSCI model contains the phrase "wedding cake"; in Figure 4h, the phrase "in front of" is used to describe the relationship between the man and fruit. The above comparison and analysis results show that the MSCI model can effectively improve the accuracy and comprehensiveness of the generated caption by iteratively updating image features and leveraging the context information between the different semantic layers. Compared with the Up-Down and SGAE models, the MSCI model can accurately describe the main objects and their relationships, the generated captions contain the mainly semantic information of the image, and the description effect is more comprehensive.

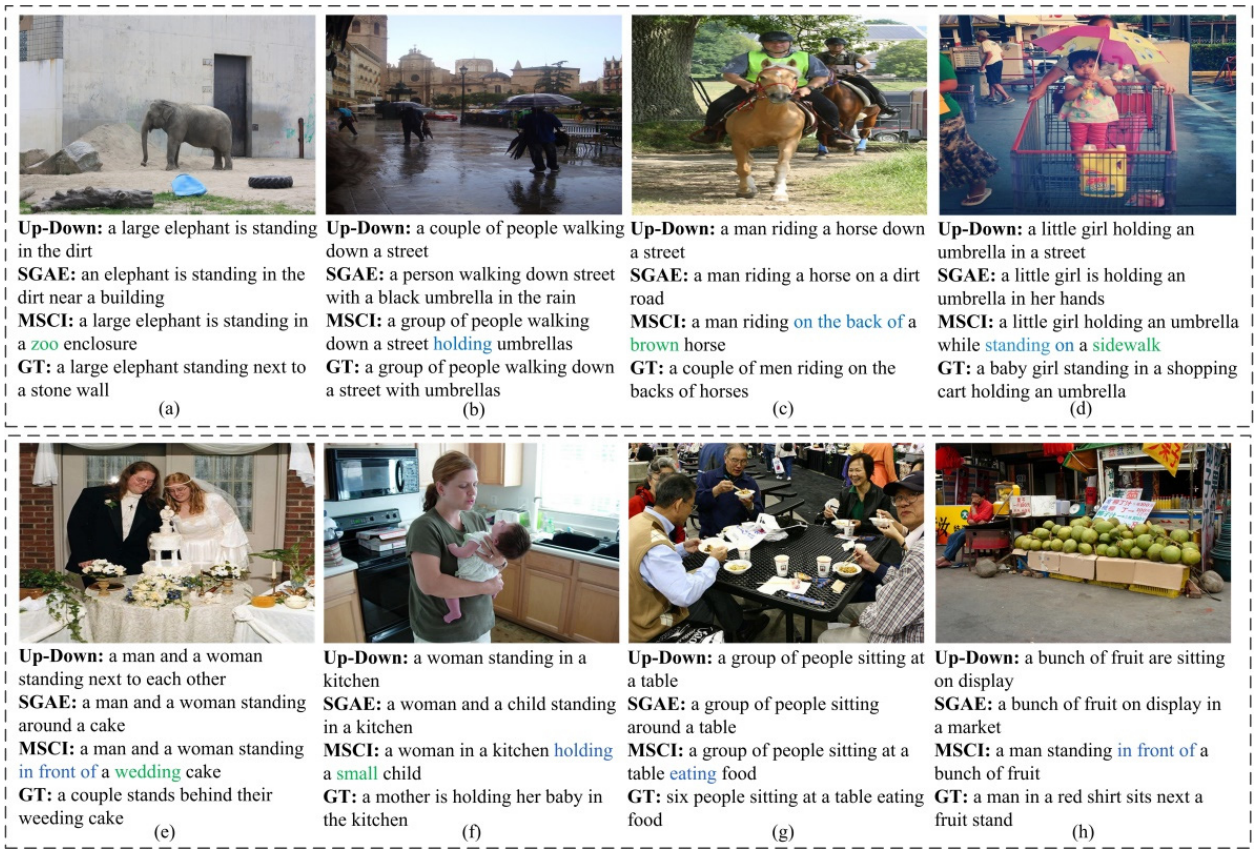

Figure 4. Examples of captions generated by the MSCI model and other models on COCO dataset and Flicker30k dataset. (**a**–**d**) display the example results of some COCO images; (**e**–**h**) display the example results of some Flicker30k images.GT denotes the ground truth reference caption.

4.4. Evaluation Analysis

4.4.1. Object Detection

As one of the fundamental visual tasks, the accuracy of object detection directly affects the generation of higher-level visual tasks. We compare our proposed MSCI model with Faster R-CNN on the COCO dataset mentioned above (4.1.1), and adopt the mean Accuracy Precision (mAP) as the evaluation metric for object detection. We checked whether the model can improve the feature refinement of the convolution layers through mutual connections and iterative updates across the features of object, relationship, and caption semantic layers, and leverage the context information between the different semantic layers

to improve the accuracy of object detection. The comparison results are demonstrated in the Table 4.

Table 4. Object detection results on the COCO dataset.

Object Det.	Faster R-CNN	Ours-CI	Ours
mAP(%)	37.4	39.5	40.8

As shown in Table 4, our model, without the context information between the different semantic layers (Ours-CI) and the mAP of object detection improvement of 2.1% than the Faster R-CNN, demonstrates that through mutual connections and iterative updates across the features of the object, relationship, and caption semantic layers can benefit the feature refining of convolution layer to improve the accuracy of object detection. Our model with the context information between the different semantic layers (Ours) and the mAP of object detection improvement of 1.3%, and improvement of 3.4% than the Faster R-CNN, demonstrates that our proposed method leverages the context information between the different semantic layers and can provide extra information for object detection; this information is helpful for object detection and recognition.

4.4.2. Relationship Detection

The accuracy of relationship detection directly affects the generation of image captioning and other higher-level visual tasks. We compare our proposed model with the existing state-of-the-art relationship detection models on the same VRD dataset. The comparison results are shown in the Table 5.

Table 5. Comparison results (in %) of our model with existing state-of-the-art models on the VRD dataset. Baseline denotes the model without the GGNN and context information. RCI denotes the relationship context information.

Model	Relationship Detection		Phrase Detection	
	R@50	R@100	R@50	R@100
LP [24]	13.86	14.70	16.17	17.03
DR-Net [25]	16.94	20.20	19.02	22.85
VRL [23]	18.19	20.79	21.37	22.60
F-Net [21]	18.32	21.20	26.03	30.77
AVR [22]	22.83	25.41	29.33	33.27
Baseline	12.58	14.29	14.26	16.32
Base+GGNN	16.85	19.32	20.65	22.28
Base+GGNN+RCI	18.36	21.27	23.58	25.56

As shown in Table 5, we can see that the scores of our model on the relationship detection and phrase detection outperformed some compared models. The experimental results demonstrate that our model can significantly improve the accuracy of vision relationship detection by refining the features of object and relationship and using the context information between object and relationship semantic layers.

To better verify the relationship detection performance of our proposed approach with different settings, we view the model without GGNN and context information as the baseline model, and perform the relationship detection in the case of with feature update and relationship context information. The results, shown at the bottom in Table 5, show that our model has a relative improvement of 5.78~9.32% than the baseline model, and with exacts more feature information and the accuracy of relationship detection gradually improve. The above discussion and comparison results show that our model can improve the accuracy of visual relationship detection by refining the object features and relationship features and leveraging the context information between the different semantic layers.

4.5. Ablation Study

The core of our proposed MSCI model is to update the features of the three different semantic layers in image, and extracting the context information between different semantic layers to improve the performance of the model-generation caption. We conduct an ablative experiment to analyze the description performance of MSCI model with different setting in Table 6, the left columns indicate whether or not we used GGNN, Relationship Context Information (RCI), Scene Context Information (SCI), and Attention Mechanism (AM) in our model, and the results are presented in the right columns.

Table 6. Ablation studies of the MSCI model on the COCO dataset.

ID	GGNN	RCI	SCI	AM	B@1	B@2	B@3	B@4	M	R	C	S
1	-	-	-	-	0.725	0.549	0.408	0.313	0.241	0.526	0.957	0.186
2	√	-	-	-	0.776	0.596	0.453	0.362	0.262	0.553	1.196	0.202
3	√	√	-	-	0.791	0.613	0.479	0.376	0.273	0.568	1.231	0.214
4	√	√	√	-	0.805	0.632	0.495	0.384	0.286	0.582	1.259	0.221
5	√	√	√	√	0.812	0.655	0.502	0.393	0.295	0.593	1.275	0.225

4.5.1. Performance Analysis

Model 1 in Table 6 is the baseline model that does not use GGNN to update the semantic features and does not have the context information and attention mechanism. Compared to Model 1, the scores of Model 2 on all metrics increased by 0.016 ~ 0.239, which shows that the model uses GGNN to iteratively update image features of three semantic layers and can leverage the connection information between objects, relationships, and caption to improve the description performance. Compared with Model 2, the scores of Model 3 and Model 4 on the evaluation metrics increased by 0.011 ~ 0.035 and 0.019 ~ 0.063, respectively. This demonstrates that exacting the relationship context and scene context between different semantic layers enables the model to obtain the richer image features to improve the accuracy and comprehensiveness of the generated caption. By comparing model 4 and model 5, we can see that the scores of Model 5 on the above evaluation metrics have further increased by 0.004 ~ 0.023. This indicates that the attention mechanism not only can guide the model to pay attention to the image features corresponding to the words when generating the caption words, but also send the image features and related context information into the decoder to generate accuracy description words to improve the accuracy and comprehensiveness of the model-generation caption.

4.5.2. Qualitative Examples

Figure 5 shows four examples of the generated captions by the MSCI model in different settings on COCO dataset. Compared with the captions generated by Model 1 in Figure 5, we can see that the captions generated by Model 2 contain the accurate object words, such as “hand”, “enclosure”, and “girl”. The captions generated by Model 3 contain the relationship information between objects, such as “on the side of”, “under”, and “next to”. The scene context information is introduced in Model 4, and the generated captions by Model 4 contain the scene information of the image, such as “street scene”, “ground”, and “zoo”. For captions generated by Model 5, they contain more semantic information, the caption sentences structure are more complex and description effect are accurate and comprehensive. For example, in Figure 5a, the caption contains more object words, such as “street”, “car”, and “light”; in Figure 5b, the caption contains the main objects and their relationships, the phrases “sitting on” and “holding” describe the relationships between “woman”, “ground”, and “umbrella”; in Figure 5c, the caption generated by the model is considered from the entire perspective of the scene image, the phrases “standing in” and “in front of ” describe the relationships between the “elephant”, “people”, and “zoo”. The above results and analysis again suggest that MSCI model iteratively updates the image features and leverages the context information between the different semantic layers to

simultaneously improve the accuracy of object detection and relationship detection and the accuracy and comprehensiveness of image captioning.

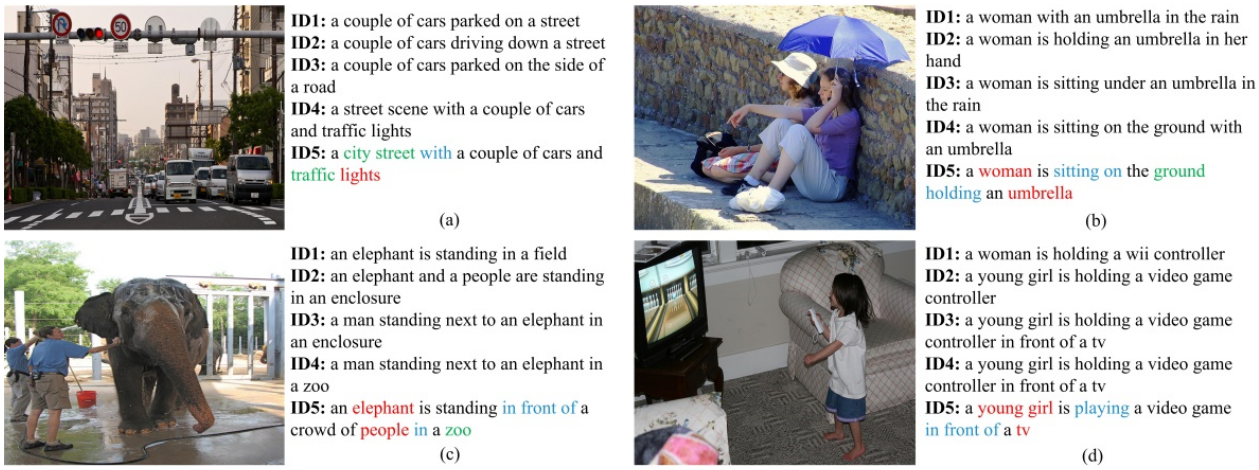

Figure 5. Qualitative examples of generated captions by our proposed MSCI model with different settings. (**a**–**d**) display the example results of some COCO images.

We further visualize the attention regions for words generated by the MSCI model in Figure 6. From the regions of attention by MSCI model, we can find out that the relationship or scene words related to the target objects in the image are generated, the model not only can adaptively attend to the image region related to the generated word at the current moment, but also focus on the related regions that have the context information with the target object. Meanwhile, the attention mechanism can send the features of these image regions and the context information related to the target object into the decoder to guide the generation of description words. For example, when predicting the word "zoo", our MSCI model not only focuses on the elephant region, but also focuses on the other related regions around the elephant.

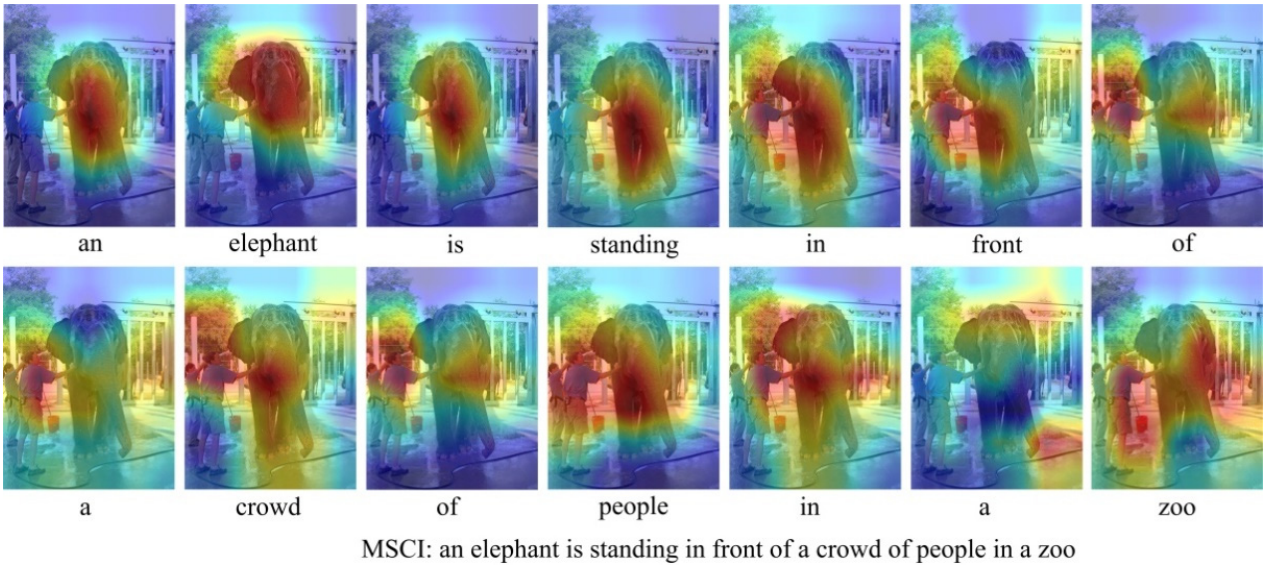

Figure 6. The visualization of attended image regions along with the caption generation process for our proposed MSCI model. For each generated word, we outline the image regions with the maximum output attribution in orange.

5. Conclusions

In this work, we propose a Multi-level Semantic Context Information (MSCI) network to jointly model object detection, relationship detection, and image-captioning tasks, which use the context information between different semantic layers in the scene image to simultaneously improve the accuracy of the three vision tasks. The model uses a feature-refining structure for mutual connections and iteratively updates the semantic features, builds a context information extraction network to extract the context information between the different semantic layers, and introduces an attention mechanism to improve the accuracy and comprehensiveness of the model-generation caption while leveraging the context information between different semantic layers to improve the accuracy of object detection, relationship detection. Experimental results on the VRD and COCO datasets demonstrate that the proposed model can simultaneously improve the accuracy of the three visual tasks and the description performance outperforms the other image-captioning models.

The method proposed in this paper is to detect and recognize the main objects in the image, reason the visual relationship between them, and generate a description corresponding to the main semantic content of the image. The accuracy of object detection and relationship detection needs to be improved, and the effect of image captioning is not satisfactory. In addition, our method solves multiple vision tasks simultaneously; the network structure is complex, and need a large amount of computer memory resources when training the model. Future work includes modifying and simplifying the network structure of the model to detect and recognize more objects, and combining with other related sensors to obtain more effective feature information to improve the accuracy of vision task generation.

Author Contributions: Methodology, P.T.; validation, P.T. and L.J.; investigation, P.T.; writing—original draft preparation, P.T.; writing—review and editing, P.T. and L.J.; visualization, P.T.; supervision, P.T. and L.J.; funding acquisition, H.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National key research and development program of China [No.2018AAA0102702] and Fundamental Research Funds for the Central Universities [No.3072020CFT0402].

Institutional Review Board Statement: This study was approved by the academic and Ethics Committee of College of Intelligent Systems Science and Engineering, Harbin Engineering University.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors are grateful for the editor and reviewers for their constructive advice on the revision of the manuscript. This work is support by the National key research and development program of China under Grant (2018AAA0102702) and Fundamental Research Funds for the Central Universities (3072020CFT0402).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. He, K.; Zhang, X.; Ren, S. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
2. Karpathy, A.; Li, F.-F. Deep Visual-Semantic Alignments for Generating Image Descriptions. In Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence, Boston, MA, USA, 7–12 June 2016; pp. 664–676.
3. Yan, M.; Guo, Y. Deep learning for visual understanding: A review. *Neurocomputing* **2016**, *187*, 27–48.
4. Wang, W.; Hu, H. Multimodal object description network for dense captioning. *IEEE Electron. Lett.* **2017**, *53*, 1041–1042. [CrossRef]
5. Johnson, J.; Karpathy, A.; Li, F.-F. DenseCap: Fully Convolutional Localization Networks for Dense Captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 4565–4574.
6. Xu, K.; Ba, J.; Kiros, R. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the 32nd International Conference on International Conference on Machine Learning (ICML), Lille, France, 6–7 July 2015; pp. 2048–2057.

7. Gu, J.; Wang, G.; Cai, J. An Empirical Study of Language CNN for Image Captioning. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1222–1231.
8. Zhang, Y.; Shi, X.; Mi, S. Image Captioning with Transformer and Knowledge Graph. *Pattern Recognit. Lett.* **2021**, *143*, 43–49. [CrossRef]
9. Zhang, Z.; Wu, Q.; Wang, Y. Exploring Region Relationships Implicitly: Image Captioning with Visual Relationship Attention. *Image Vis. Comput.* **2021**, *109*, 104146. [CrossRef]
10. Zhou, Y.; Sun, Y.; Honavar, V. Improving Image Captioning by Leveraging Knowledge Graphs. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa Village, HI, USA, 8–10 January 2019; pp. 283–293.
11. You, Q.; Jin, H.; Wang, Z. Image captioning with semantic attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 4651–4659.
12. Lu, J.; Xiong, C.; Parikh, D. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 375–383.
13. Gao, L.; Fan, K.; Song, J. Deliberate Attention Networks for Image Captioning. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Honolulu, HI, USA, 27–31 January 2019; pp. 33, 8320–8327.
14. Yang, X.; Tang, K.; Zhang, H. Auto-Encoding Scene Graphs for Image Captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–21 June 2019; pp. 10685–10694.
15. Zhong, Y.; Wang, L.; Chen, J. Comprehensive Image Captioning via Scene Graph Decomposition. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 211–229.
16. Li, Y.; Tarlow, D.; Brockschmidt, M. Gated Graph Sequence Neural Networks. In Proceedings of the IEEE International Conference on Learning Representations (ICLR), San Juan, PR, USA, 2–4 May 2016; pp. 1–20.
17. Girshick, R.; Donahue, J.; Darrell, T. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
18. Ren, S.; He, K.; Girshick, R. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
19. Bochkovskiy, A.; Wang, C.-Y.; Liao, H. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934v1.
20. Liu, W.; Anguelov, D.; Erhan, D. SSD: Single Shot MultiBox Detector. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
21. Li, Y.; Ouyang, W.; Zhou, B. Factorizable net: An efficient subgraph-based framework for scene graph generation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 335–351.
22. Lv, J.; Xiao, Q.; Zhong, J. AVR: Attention based Salient Visual Relationship Detection. *arXiv* **2020**, arXiv:2003.07012.
23. Liang, X.; Lee, L.; Xing, E.P. Deep Variation-structured Reinforcement Learning for Visual Relationship and Attribute Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 848–857.
24. Lu, C.; Krishna, R.; Bernstein, M. Visual Relationship Detection with Language Priors. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 852–869.
25. Dai, B.; Zhang, Y.; Lin, D. Detecting visual relationships with deep relational networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3076–3086.
26. Chen, T.; Yu, W.; Chen, R. Knowledge-Embedded Routing Network for Scene Graph Generation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 6156–6164.
27. Kulkarni, G.; Premraj, V.; Ordonez, V. Babytalk: Understanding and generating simple image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 2891–2903. [CrossRef] [PubMed]
28. Elliott, D.; Keller, F. Image description using visual dependency representations. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP), Seattle, WA, USA, 18–21 October 2013; pp. 1292–1302.
29. Verma, Y.; Gupta, A.; Mannem, P. Generating image descriptions using semantic similarities in the output space. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 23–28 June 2013; pp. 288–293.
30. Devlin, J.; Cheng, H.; Fang, H. Language models for image captioning: The quirks and what works. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), Beijing, China, 26–31 July 2015; pp. 100–105.
31. Zhu, Z.; Wang, T.; Qu, H. Macroscopic Control of Text Generation for Image Captioning. *arXiv* **2021**, arXiv:2101.08000.
32. Ji, J.; Xu, C.; Zhang, X. Spatio-Temporal Memory Attention for Image Captioning. *IEEE Trans. Image Process.* **2020**, *29*, 7615–7628. [CrossRef]
33. Anderson, P.; He, X.; Buehler, C. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake, UT, USA, 18–23 June 2018; pp. 6077–6086.
34. Wang, W.; Chen, Z.; Hu, H. Hierarchical Attention Network for Image Captioning. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Honolulu, HI, USA, 27–31 January 2019; pp. 8957–8964.
35. Mi, J.; Lyu, J.; Tang, S. Interactive Natural Language Grounding via Referring Expression Comprehension and Scene Graph Parsing. *Front. Neurorobot.* **2020**, *14*, 43. [CrossRef] [PubMed]

36. Li, X.; Jiang, S. Know More Say Less: Image Captioning Based on Scene Graphs. *IEEE Trans. Multimed.* **2019**, *21*, 2117–2130. [CrossRef]
37. Mottaghi, R.; Chen, X.; Liu, X. The Role of Context for Object Detection and Semantic Segmentation in the Wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 891–898.
38. Zeng, X.; Ouyang, W.; Yang, B. Gated Bi-directional CNN for Object Detection. In Proceedings of the IEEE European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 354–369.
39. Ma, Y.; Guo, Y.; Liu, H. Global Context Reasoning for Semantic Segmentation of 3D Point Clouds. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Snowmass Village, CO, USA, 2–5 March 2020; pp. 2931–2940.
40. Lin, C.-Y.; Chiu, Y.-C.; Ng, H.-F. Global-and-Local Context Network for Semantic Segmentation of Street View Images. *Sensors* **2020**, *20*, 2907. [CrossRef] [PubMed]
41. Dvornik, N.; Mairal, J.; Schmid, C. On the Importance of Visual Context for Data Augmentation in Scene Understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 2014–2018. [CrossRef] [PubMed]
42. Zhuang, B.; Liu, L.; Shen, C. Towards Context-Aware Interaction Recognition for Visual Relationship Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 589–598.
43. Zellers, R.; Yatskar, M.; Thomson, S. Neural motifs: Scene graph parsing with global context. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake, UT, USA, 18–23 June 2018; pp. 5831–5840.
44. Qi, X.; Liao, R.; Jia, J. 3D Graph Neural Networks for RGBD Semantic Segmentation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
45. Kenneth, M.; Ruslan, S.; Abhinav, G. The More You Know: Using Knowledge Graphs for Image Classification. *arXiv* **2017**, arXiv:1612.04844.
46. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
47. Lin, T.-Y.; Maire, M.; Belongie, S. Microsoft coco: Common objects in context. In Proceedings of the IEEE European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
48. Plummer, B.-A.; Wang, L.; Cervantes, C.-M. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 13–16 December 2015; pp. 2641–2649.
49. Papineni, K.; Roukos, S.; Ward, T. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL), Philadelphia, PA, USA, 6–12 July 2002; pp. 311–318.
50. Banerjee, S.; Lavie, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, MI, USA, 29 June 2005; pp. 65–72.
51. Lin, C.-Y.; Hovy, E. Automatic evaluation of summaries using n-gram co-occurrence statistics. In Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), Baltimore, MD, USA, 1–11 July 2003; pp. 71–78.
52. Vedantam, R.; Zitnick, C.; Parikh, D. Cider: Consensus-based image description evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 4566–4575.
53. Anderson, P.; Fernando, B.; Johnson, M. Spice: Semantic propositional image caption evaluation. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 382–398.
54. Kingma, D.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
55. He, X.; Shi, B.; Bai, X. Image caption generation with part of speech guidance. *Pattern Recognit. Lett.* **2019**, *119*, 229–237. [CrossRef]
56. Nogueira, T.; Vinhal, C. Reference-based model using multimodal gated recurrent units for image captioning. *Multimed. Tools Appl.* **2020**, *79*, 30615–30635. [CrossRef]