Building Joint Relationship Attention Network for Image-Text Generation

Changzhi Wang and XiaoDong Gu*

Department of Electronic Engineering, Fudan University
Shanghai, 200438, China
18110720047@fudan.edu.cn(C. Wang) xdgu@fudan.edu.cn(X. Gu)

Abstract

Attention based methods for image-text generation often focus on visual features individually, while ignoring relationship information among image features that provides important guidance for generating sentences. To alleviate this issue, in this work we propose the Joint Relationship Attention Network (JRAN) that novelly explores the relationships among the features. Specifically, different from the previous relationship based approaches that only explore the single relationship in the image, our JRAN can effectively learn two relationships, the visual relationships among region features and the visual-semantic relationships between region features and semantic features, and further make a dynamic trade-off between them during outputting the relationship representation. Moreover, we devise a new relationship based attention, which can adaptively focus on the output relationship representation when predicting different words. Extensive experiments on large-scale MSCOCO and small-scale Flickr30k datasets show that JRAN achieves state-of-the-art performance. More remarkably, JRAN achieves new 28.3% and 58.2% performance in terms of BLEU4 and CIDEr metric on Flickr30k dataset.

1 Introduction

Image-text generation (i.e., image captioning) is a typical cross-modal task that connects Natural Language Processing (NLP) and Computer Vision (CV) (Tahvili et al., 2020). Its core goal is to automatically predict a meaningful and grammatically correct sentence, which can accurately describe the main content of images. Practical applications for this task mainly include injecting visual intelligence into the chatbots, searching semantic image, and helping people with visual impairments to understand the visual world. However, image-text generation is still a challenging task. The main difficulties originate from two aspect: **1)** The noise and complex background in the image are likely to interfere with the generation of correct caption; **2)** The interaction between features in the image is often overlooked.

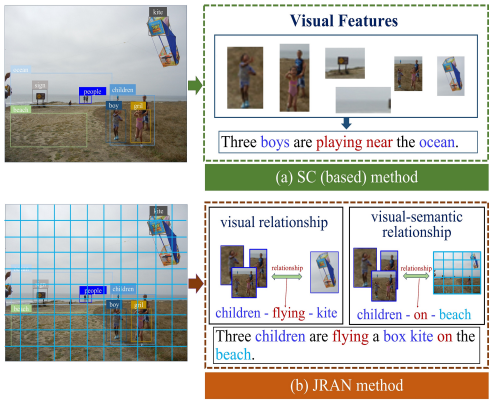


Figure 1: Illustration of different schemes. (a) is an example predicted by SC (base) baselines (Chen et al., 2017) that mainly uses visual features to generate sentence. (b) presents a more accurate sentence generated by our JRAN method which learns two relationship, i.e., visual relationship between region features and visual-semantic relationship between region features and semantic features (i.e., background and environment).

For difficulty **1)**, encouraging by the method (Wu et al., 2021b), the noise can be injected into RNN hidden states to predict the mean and standard deviation, and manipulate the RNN transition states. In this way, the network robustness can be significantly enhanced and the issue can be well solved. However for difficulty **2)**, although some related visual attention based methods (Xu et al., 2015; Wang et al., 2016; Song et al., 2018) achieve remarkable progress, they usually focus on the image visual features while ignoring the relationships between them. This makes the model often difficult to generate an accurate or appropriate description that can correctly describe the relationships among objects in the image. For example, as illustrated in Figure 1, only using the detected visual features in

*Corresponding author.

Proceedings of the 29th International Conference on Computational Linguistics, pages 5521–5531
October 12–17, 2022.

the image, SC (base) method (Chen et al., 2017) predicts a description "*Three boys are playing near the ocean.*", where the verb phrase "playing near" indicates the relationship between the objects "boy" and "ocean". Obviously, this description cannot accurately reflect the main scene of the image. Contrarily, our JRAN accurately describes the main content of image by effectively learn the visual relationship between object region features and the visual-semantic relationship between region features and semantic features, generating a more relevant sentence "*Three children are flying a box kite on the beach.*" The word "flying" appropriately describes the relationship between the two region features "children" and "kite", and the word "on" accurately represents the relationship between region information "children" and semantic information "beach" (i.e., background/surrounding). Thus, learning the relationship between image features is of crucial importance for generating accurate sentence description for image.

Based on the above observations, different from previous relationship based approaches (Wang et al., 2020; Kipf and Welling, 2017; Li and Jiang, 2020) (See Figure 2(top)) that only explore the single feature relationship in the image, we present a new Joint Relationship Attention Network (JRAN) that novelly learns the joint relationship between region features and semantic features in Figure 2(bottom).

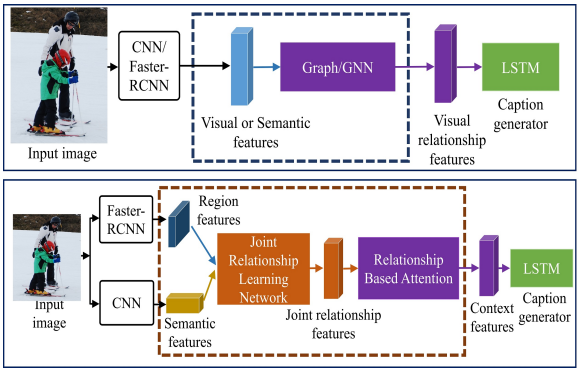


Figure 2: The illustration of existing relationship based methods (top) and our JRAN method (bottom). (a) Relationship based methods usually use Graph/GNN (Graph Neural Network) to explore single (visual or semantic) feature relationship; (b) Our JRAN method can learn the joint relationship between visual features and semantic features.

Specifically, we first utilize the object detector Faster R-CNN and CNN to extract region features and complementary semantic features from the input image. Then, JRAN builds two types of relationship, i.e., visual relationship and visual-semantic relationship. The former is solely based on the detected region features in the image, and encodes the visual relationship between region features. Instead, the latter takes the region features and semantic features of the image into account to fully explore the visual-semantic relationship between them. As shown in Figure 3, to obtain the image representation containing joint relationship features, we devise a core competition module called joint relationship learning network, which can effectively learn the visual relationship and visual-semantic relationship while dynamically balancing the different contributions between them. After that, we introduce the relationship based attention module, which can adaptively focus on the obtained most relevant joint relationship features during generating words. The whole framework of JRAN could be jointly learnt and optimized in an end-to-end way. Our JRAN method achieves significant improvement compared with the related relationship based methods. More remarkably, it can be integrated into a better baseline model to achieve better performance.

Our contributions mainly include: firstly, we propose a novel image-text generation network that utilizes the complementary region and semantic features in the image for enriching feature representations. Secondly, we devise a joint relationship learning network, which can fully learn both visual relationship and visual-semantic relationship in the image, and further balance their contributions during predicting different words. Finally, exhaustive experiments indicate that our JRAN is not only effective on large-scale but also achieves superior performance on small-scale datasets.

2 Related Works

2.1 Image-Text Generation

Image-text generation can be treated as a sequence-to-sequence task, which converts the data from a raw image to a sentence description. For example, Wang et al. (2016) presented an end-to-end architecture to generate the sentence where visual embedding is encoded with CNN and sentence embedding is encoded using Bi-LSTM. Xu et al. (2015) presented the first model based on visual attention, where it extracts visual features of each region from the raw image, and then assigns different weights for them. Chen et al. (2017) in-

tegrated the spatial and channel attention features extracted from a CNN, and uses the channel attention to focus on different semantic information. Similarly, Song et al. (2018) integrated the spatial and channel attentions into salient object regions, and effectively improved the performance of Visual Question Answering (VQA) task. Although the above approaches have well performance, they ignore the relationships among image features. In our work, on one hand, we take full advantage of the two complementary region features and semantic features in the image. On the other hand, we learn the visual relationships among region features while exploring the visual-semantic relationships between region features and semantic features.

2.2 Relationship Based Approaches

Recently, relationship based methods have been proposed to boost the performance for image-text generation task. It mainly uses Graph Convolution Network (GCN) and Graph Attention neTwork (GAT) to learn the single relationship between local features or global features. For instances, Li et al. (2018b) took the data of arbitrary graphic structure as the input and introduced a flexible and general GCN. Kipf and Welling (2017) presented the GCN, which can be directly used to process graph structure data. Wang et al. (2020) utilized GAT to learn the relationship between image features. It directly inputs the extracted region features and semantic features into the GAT, and then follows the self-attention strategy to calculate the relationship between each feature node. Different from previous relationship based methods, we devise a new joint relationship attention network, which can capture the visual relationship and visual-semantic relationship in the image, and then further balance their different contributions during generating different words. This effectively promotes the model performance.

3 Our approach

The main purpose of this work is to explore the relationship between different features in the image, so as to generate a sentence description containing accurate interaction information for the input image. The overall architecture of the proposed JRAN is shown in Figure 3. The critical elements in the architecture are described in detail as follows.

3.1 Problem Formulation

Formally, a sentence model receives a source image I as the input and is required to output a target text sentence S to describe the image main content. S is a sequence of sentence generated word by word, which can be presented as $S = \{x_1, x_2.., x_T\}$, where T denotes the length of the sequence, and x_t, $t \in [1, T]$ is the t-th word. During the training, given a training dataset with a set of image-sentence pairs (I_i, S_i), and sentence model is trained to minimize the cross entropy loss which is equivalent to maximizing the likelihood,

$$L_{loss}(\theta) = -\sum_{i=1}^{M}\sum_{t=1}^{T}(logp(x_{i,t}|I_i, x_{i,1:t-1}, \theta)), \quad (1)$$

where θ is the model parameters needed to train, M is the total number of training samples, and $x_{i,t}$ denotes the t word of ground-truth caption S_i.

3.2 Feature Extraction

We extract the two complementary features from the raw image: region features $\mathbf{V}_r$ and semantic features $\mathbf{V}_s$. For region features $\mathbf{V}_r \in \mathbb{R}^{D\times K}$, the raw image is first fed into Faster R-CNN to detect the top K candidate visual regions. For each selected region k, we take the mean-pooled convolutional feature from the image region as $\mathbf{V}_r^k$, which has D dimensions. Thus, the region features $\mathbf{V}_r = [\mathbf{V}_r^1, ..., \mathbf{V}_r^K]$, $\mathbf{V}_r^k \in \mathbb{R}^D$. For semantic features $\mathbf{V}_s \in \mathbb{R}^{L\times D}$, since last convolutional layer ($Conv5\_3$) of ResNet usually contains the context (or background) information around objects (Li et al., 2018a), thus, it is extracted as the image semantic features. Then, the extracted feature map $\mathbf{V}_l \in \mathbb{R}^{W\times H\times D}$ is further flattened into $\mathbf{V}_s \in \mathbb{R}^{L\times D}$, $L = W \times H$,

$$\mathbf{V}_s = \{\mathbf{V}_s^1, ..., \mathbf{V}_s^D\} = flatten(Conv(I)), \quad (2)$$

where $\mathbf{V}_s^i \in \mathbb{R}^L$,$i \in \{1, 2, .., D\}$ represents the i-th semantic feature of the feature map $\mathbf{V}_s$.

3.3 Joint Relationship Learning Network

We devise the Joint Relationship Learning Network (JRLN) in Figure 4. It is consists of some stacked feature relationship network, and each feature relationship network is composed of a multi-head Relationship Computation (RC) module. On one hand, JRLN learns the visual relationship between region features, which can unfold the inherent action/interaction between different region objects.

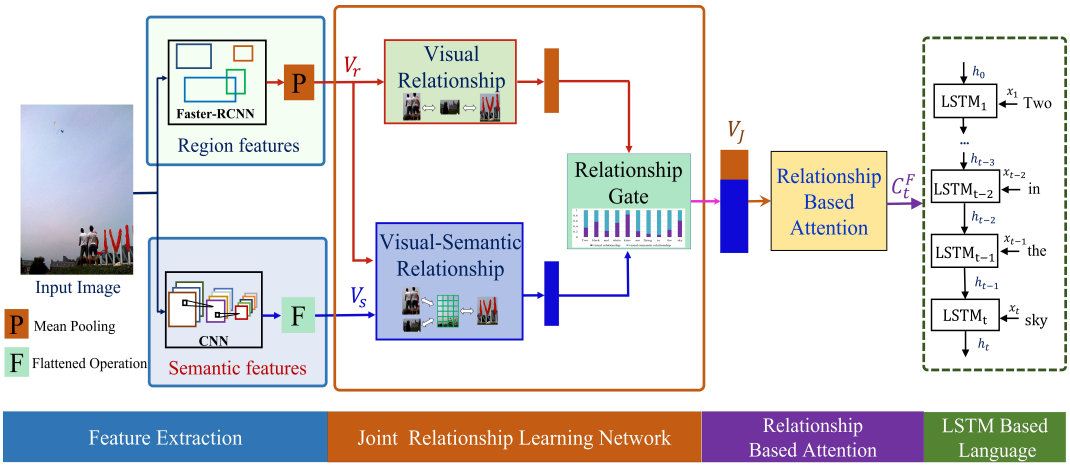


Figure 3: An overall framework of our proposed JRAN which consists of four modules: Feature Extractor module, Joint Relationship Learning Network, Relationship Based Attention module and LSTM Based Language module. We first use Feature Extractor module to extract region features V_r and complementary semantic features V_s for image feature representation. Then, these features are input into the joint relationship learning network respectively, and the balanced relationship feature V_J are output. After that, V_J is fed into relationship based attention module to obtain the final context representation C_t^F. Finally, C_t^F is input into the LSTM Based Language module for word generation.

On the other hand, JRLN utilizes the image semantic features as guide to learn the visual-semantic relationship between the semantic features and complementary region features, which effectively connects isolated region objects with their background (or environment) information. Importantly, JRLN further balances the different contributions between the two relationships during generating words.

Now, we describe our competitive joint relationship learning network. It consists of some Relationship Attention (RA) and Feed-Forward Network (FFN) modules, in which RA includes N_r RC modules. And the RC is able to learn two kinds of relationships: *(I)* Visual relationships among region features, *(II)* Visual-semantic relationships between region features and semantic features.

I) Visual relationships among region features: Considering that the region features in the image do not exist independently of each other, we learn the visual relationship among region features by using a multi-head RC, as illustrated in Figure 4 (b)-(I). Given the input set of N region features $\{\mathbf{V}_r^i\}$, the output visual relationship feature $\mathbf{V}_v^i$ of the whole region feature with respect to the i-th region feature is calculated as follows

$$\mathbf{V}_v^i = \sum_j \varphi^{ij} \cdot (\mathbf{W}_v \mathbf{V}_r^j), \tag{3}$$

where $\mathbf{W}_v$ is the model learnable matrix. Further, the visual relationship weight φ^{ij} is calculated by measuring the correlation between the i-th region feature and the j-th region feature,

$$\varphi^{ij} = softmax(\frac{\mathbf{W}_q \mathbf{V}_r^i \cdot (\mathbf{W}_k \mathbf{V}_r^j)^T}{\sqrt{d_k}}), \tag{4}$$

where $\mathbf{W}_q$ and $\mathbf{W}_k$ are the projection matrices of the i-th and j-th region features. d_k is the matrix dimension after projection (i.e., scaling factor). Eq. (4) reflects how much every region is affected by other regions, where semantically more corresponding regions may have higher relationship weight values in the image.

II) Visual-semantic relationships between region features and semantic features: Since region features and semantic features are related to some extent, how to learn the relationship between them is important for image-text generation. Generally, to effectively organize the region features guided by the image semantic features, a common and straightforward idea is to concatenate the guided semantic features and all region features (i.e., Figure 5(b)). However, such a scheme is too naive to model the relative importance of the region and semantic features, i.e., the discrimination introduced by semantic guiding features and region features are different and should be distinguished. Thus, we propose to utilize an attention mechanism to weight the relative relationship between region features and semantic features, as illustrated in Figure 4 (d)-(II). Given the input region features $\mathbf{V}_r$ and

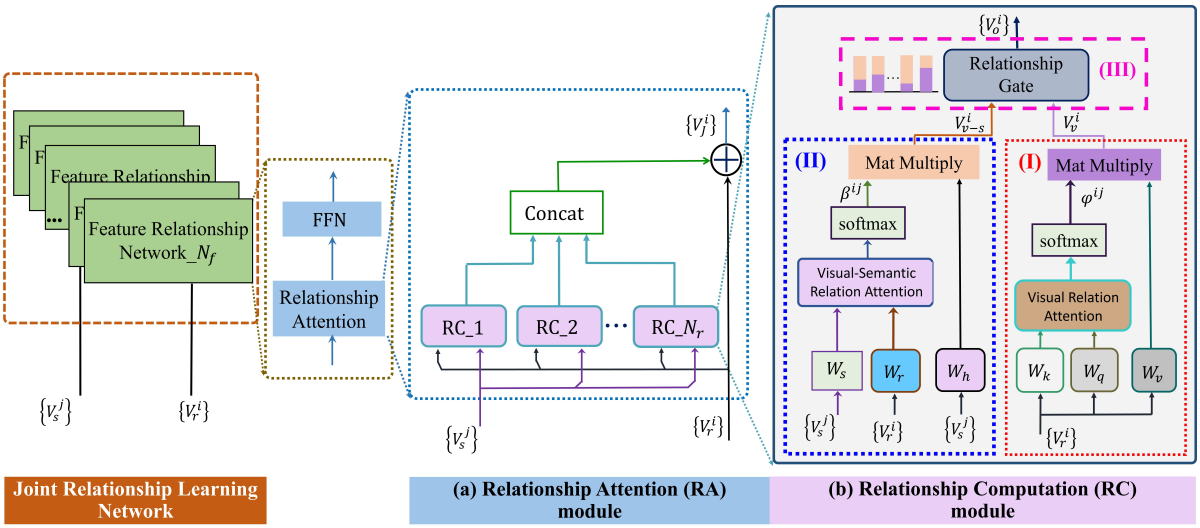


Figure 4: The illustration of our Joint Relationship Learning Network. It consists of a stack of Feature Relationship Network, and each feature relationship network includes a Feed-Forward Network (FFN) and a Relationship Attention (RA) module (a). Further, each RA consists of N_r Relation Computation (RC) modules (b). In addition, in RC module (b), the broken red line part (I) indicates the visual relationships among region features; the broken blue line part (II) is the visual-semantic relationships between region features and semantic features, and the broken pink line part (III) denotes our designed Relationship Gate.

semantic features $\mathbf{V}_s$ of the raw image, the output visual-semantic relationship feature $\mathbf{V}^i_{v-s}$ is,

$$\mathbf{V}^i_{v-s} = \sum_j \beta^{ij} \cdot (\mathbf{W}_h \mathbf{V}^j_s), \tag{5}$$

where $\mathbf{V}^i_{v-s}$ denotes the i-th visual-semantic relationship feature between region features and semantic features. Further, the visual-semantic relationship weight β^{ij} between the i-th region feature and j-th semantic feature is computed according to

$$\beta^{ij} = softmax(\frac{\mathbf{W}_s \mathbf{V}^j_s \cdot (\mathbf{W}_r \mathbf{V}^i_r)^T}{\sqrt{d_k}}), \tag{6}$$

where β^{ij} reflects the influence of image semantic features on region features. $\mathbf{W}_s$ and $\mathbf{W}_r$ are the model learnable matrices, which project the original semantic features $\mathbf{V}^j_s$ and region features $\mathbf{V}^i_r$ into the subspaces to measure how well they match, and d_k is the feature dimension after projection.

III) Relationship gate: As described in Section introduction, relationships are of crucial important for accurately describing the main content of the input images. Considering that visual relationships $\mathbf{V}^i_v$ and visual-semantic relationships $\mathbf{V}^i_{v-s}$ play different roles during generating different words. Thus, we introduce a relationship gate, which can dynamically balance their different contributions to obtain the image feature representation $\mathbf{V}^i_o$ containing different relationship information,

$$\mathbf{V}^i_o = \sigma \cdot \mathbf{W}_{vs} \mathbf{V}^i_v + (1 - \sigma) \cdot \mathbf{W}_{vb} \mathbf{V}^i_{v-s}, \tag{7}$$

where σ is the relationship gate coefficient, as,

$$\sigma = sigmoid(\text{Concat}(\mathbf{U}_{vr} \mathbf{V}^i_r, \mathbf{U}_{vs} \mathbf{V}^i_s) + \mathbf{b}_\sigma), \tag{8}$$

where $\mathbf{W}$ and $\mathbf{U}$ are the learnable matrixes, and $\mathbf{b}_\sigma$ is a bias.

Further, to comprehensively learn visual relationship while capturing visual-semantic relationship, in our model we devise the multi-head RC in which each head can focus on different relationship attributes. Specifically, the relationship attention module aggregates in total N_r RC modules,

$$\mathbf{V}^i_J = \mathbf{S}_r \mathbf{V}^i_r + \text{Concat}(\mathbf{V}^i_o(1), ..., \mathbf{V}^i_o(N_r)) \mathbf{S}_n, \tag{9}$$

where $\mathbf{S}_r$ is learnable parameter, and $\mathbf{S}_n$ is the output projection matrix that aggregates the information from different heads, and $\mathbf{V}^i_o(n)$, $n \in (1, N_r)$ denotes the n-th relationship feature.

Finally, a basic feed-forward network is complemented to increase the model non-linearity, which takes V^i_J as its input and outputs as follows

$$\mathbf{V}^i_J = \mathbf{V}^i_J \mathbf{S}_J + \mathbf{b}_J. \tag{10}$$

where $\mathbf{S}_J$ and $\mathbf{b}_J$ are the learnable parameters. $\mathbf{V}^i_J$ denotes the i-th joint relationship feature obtained by relationship attention module.

3.4 Relationship Based Attention

Although the obtained joint relationship feature $\mathbf{V}_J^i$ have provided a full relationship representation for the image, in many cases, a word/phrase in the generated sentence is only related to some of the specific information containing in the feature representation. Thus, we further develop a relationship based attention module to automatically attend to the corresponding relationship feature during generating words. The final attention context vector C_t^F is obtained by the following updates,

$$z_{i,t} = \mathbf{W}_r^T tanh(\mathbf{W}_J \mathbf{V}_J^i + \mathbf{W}_h \mathbf{h}_{t-1} + \mathbf{b}_h), \quad (11)$$

$$\eta_t = softmax(\mathbf{z}_t), \quad (12)$$

$$\mathbf{C}_t^F = \sum \eta_t \mathbf{V}_J^i, \quad (13)$$

where $\mathbf{W}$ and $\mathbf{b}_h$ are the model learnable parameters. Next, $\mathbf{C}_t^F$ will be fed into the LSTM based language module to predict word.

4 Experiments

4.1 Experiment Setting

Datasets and Evaluation Metrics: We extend the experiment from large-scale MSCOCO (Lin et al., 2014) to small-scale Flickr30k (Plummer et al., 2015) datasets to verify the effectiveness of our model. Further, several popular evaluation metrics: BLEU (Papineni et al., 2002), ROUGE-L (R) (Lin, 2004), METEOR (M) (Banerjee and Lavie, 2005), CIDEr (C) (Vedantam et al., 2015) and SPICE (S) (Anderson et al., 2016) are used to evaluate the model performance, and coco-caption code[1] is utilized to compute these metrics.

Implementation Details: The LSTM hidden state dimension is set to 512, and the number of hidden cells and the embedded size of input words are also set to 512. Further, the bottom-up features provided by UD (Anderson et al., 2018) is also used. We use the gradient clipping strategy during back propagation to alleviate the problem of gradients explosion. The initial learning rate of CNN is 1e-5 and that of language model is 5e-4. When fine-tuning the image model, the learning rate we used is considerably smaller than that originally used for the training model. In 24 training epochs, the model stops training if its performance is not improved. In addition, these experiments are implemented via PyTorch, and we use Beam Search (BS) strategy for predicting caption.

[1] Available: https://github.com/tylin/coco-caption

4.2 Ablation Studies

Firstly, some ablation experiments are performed to clarify the effectiveness of following modules: 1) Region Features (R Fea.), 2) Semantic Features (S Fea.), 3) Joint Relationship Learning Network (JRLN), 4) Relationship based Attention (R-Att.) module. Then, the effects of different relationship fusion schemes are further analyzed in detail.

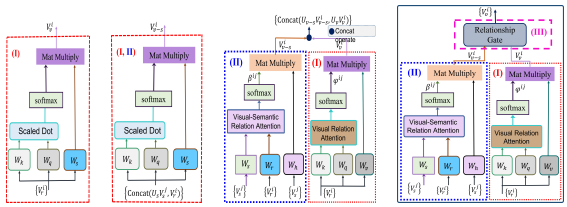
a. Effectiveness of Each Module: As shown in Table 1, **1)** in lines 1 and 2, "R Fea." or "S Fea.", "JRLN" means that model only uses the separate region features or semantic features to build the relationship; **2)** "R Fea." and "S Fea." means that region features and semantic features are directly concatenated, and then directly input into LSTM to generate sentence; **3)** "R Fea.", "S Fea." and "JRLN" means that model learns visual relationship and visual-semantic relationship, but it does not introduce relationship attention to focus on them; **4)** "R Fea.", "S Fea." and "R-Att" denotes that a relationship attention module is directly used to focus on the concatenated region and semantic features; **5)** in line 6, "R Fea.", "S Fea.", "JRLN" and "R-Att" is our full model, which explores the two relationships among image features, and then exploits relationship attention to dynamically focus on the obtained relationship representation.

Num.	Model Settings				Model Metrics				
	R Fea.	S Fea.	JRLN	R-Att.	B-1	B-4	M	R	C
1	✓		✓		77.9	36.1	26.2	56.3	120.8
2		✓	✓		77.4	35.6	25.7	56.0	120.6
3	✓	✓			78.8	37.1	26.8	56.8	121.4
4	✓	✓	✓		80.6	38.2	28.1	58.1	127.9
5	✓	✓		✓	79.1	37.4	27.1	57.1	123.2
6	✓	✓	✓	✓	**81.0**	**38.6**	**28.3**	**58.3**	**128.4**

Table 1: Ablation performance of JRAN model on MSCOCO dataset. '✓' means that the model only uses the module for image-text generation.

We have the following conclusions from Table 1: **1)** The metric scores line 4 is higher than line 3, which shows that the designed JRLN can effectively learn the visual-semantic relationship between region features and semantic features. **2)** Line 4 outperforms separate lines 2 and 1, it indicates that visual relationship and visual-semantic relationship can complement each other. **3)** The performance of model line 6 is better than line 4, it indicates that relationship based attention module can boost the model performance. **4)** Our full model in line 6 obtains the best score, which demonstrates the overall effectiveness of the proposed model.

b. Effectiveness of Relationship Fusion Scheme:



(a) *None* (b) *Concat* (c) *No-gate* (d) *RC (Ours)*

Figure 5: Different schemes for exploring the relationship between image features. (a) *None* and (b) *Concat* utilize *scaled dot-product attention*; (c) *No-gate* directly concatenates the two relationships. (d) *RC (Ours)* is our designed relationship computation (RC) scheme.

There are generally four directions for deeply exploring the relationship among features in the image, as shown in Figure 5. **(a)** *None* denotes that model only learns the visual relationship among region features by using *scaled dot-product attention* model (Vaswani et al., 2017). Similarly, **(b)** *Concat* follows the recent methods (Vaswani et al., 2017; Duan et al., 2017), i.e., region features and semantic features are directly concatenated, and then directly fed into *scaled dot-product attention* model to learn the visual-semantic relationships. **(c)** *No-gate* does not use our designed relationship gate, ignoring the different contributions between visual relationships and visual-semantic relationships during generating different words. **(d)** *RC (Ours)* is our full competitive RC module, which fully learns the two types of relationships, and further takes into account that different relationships contribute different during generating words.

We compare the performances of the RC variants in the four schemes. The results are 36.3%, 37.9%, 38.4% and 38.6% in BLUE4 metric and 120.9%, 122.2%, 128.1% and 128.4% in CIDEr metric for (a), (b), (c) and (d) schemes, respectively. It indicates that our designed RC scheme outperforms other learning relationship schemes.

5 Experimental Results

5.1 Comparison with State-of-The-Arts

Note that, for fair comparison, these models (Cornia et al., 2020; Yao et al., 2019) are not included in the comparison since the former (Cornia et al., 2020) uses the extra dataset nocaps (Kuznetsova et al., 2018), and the latter (Yao et al., 2019) utilizes the extra COCO-detect to segment the whole object.

a. Results on MSCOCO Dataset: The comparison results on MSCOCO are shown in the left part of Table 2. It can be observed that our JRAN achieves promising results. Compared with the typical baselines (Li et al., 2018a), our newly proposed JRAN can significantly improve BLEU 4 score from 36.4 to 38.6 (6.04%) and CIDEr score from 122.2 to 128.4 (5.07%), which is a significant improvement.

b. Results on Flickr30k Dataset: To further evaluate the model generalization ability, we conduct experiments on small-scale Flickr30k dataset. As shown in the right part of Table 3, JRAN achieves the superior performance. This further demonstrates that our model still maintains good generalization ability even on small-scale dataset.

5.2 Comparison with Similar Relationship Based Methods

More importantly, relationship based methods also explore the visual relationships among image features by using the graph/GCN. For example, method "KMSL" (Li and Jiang, 2020) explicitly utilizes the semantic relationship triples (scene graph) as additional inputs to explores the visual relationship. Similarly, "ARL" (Wang et al., 2020) explores the visual relationship among image regions by using GNN/GCN. As showed in Table 3, the performance of our JRAN is significantly better than these relationship based methods across all metrics, which quantitatively demonstrates the potentials of our joint relationship attention network.

5.3 Comparison with Transformer Based Baselines

In particular, to demonstrate that our JRAN can be integrated into the current mainstream transformer based baselines to achieve a better performance, we upgraded our baselines with a plain model "Simplistic Transformer architecture (*Sim-Trans*)[2]" as new baselines. The model doesn't use transformer encoder and the projected visual features are directly processed by the transformer decoder. Therefore, to see the real performance gain contributed by our JRAN model, we feed the final attention context features C_t^F from the relationship based attention module into the transformer decoder. In Table 4, the last few rows show the results of our

[2]https://github.com/krasserm/fairseq-image-captioning

Methods	MSCOCO								Flickr30k						
	B-1	B-2	B-3	B-4	M	R	C	S	B-1	B-2	B-3	B-4	M	R	C
ALT-ALTM (Ye et al., 2018)	75.1	59.0	45.7	35.5	27.4	55.9	110.7	20.3	68.5	50.7	37.0	27.0	21.2	48.0	56.2
SCST (Gao et al., 2019)	77.9	61.5	46.8	35.0	26.9	56.3	115.2	20.42	-	-	-	-	-	-	-
VD-SAN (He et al., 2019)	73.4	56.6	42.8	32.2	25.4	-	99.9	-	65.2	47.1	33.6	23.9	19.9	-	-
Up-Down (Anderson et al., 2018)	79.8	-	-	36.3	27.7	56.9	120.1	21.4	-	-	-	-	-	-	-
GLA (Li et al., 2018a)	72.5	55.6	41.7	31.2	24.9	53.3	96.4	-	56.8	37.2	23.2	14.6	16.6	41.9	36.2
HAN (Wang et al., 2019)	80.9	64.6	**49.8**	37.6	27.8	58.1	121.7	21.5	-	-	-	-	-	-	-
Trans+KG (Zhang et al., 2021)	76.24	-	-	34.39	27.71	-	112.60	21.12	68.36	-	-	26.55	21.71	-	56.62
TDA+GLD (Wu et al., 2021a)	78.8	62.6	48.0	36.1	27.8	57.1	121.1	21.6	-	-	-	-	-	-	-
cLSTM-RA (Yang et al., 2020)	**81.7**	64.5	49.4	37.2	28.0	57.9	121.5	-	70.5	52.5	37.6	27.1	21.9	49.4	57.7
Baselines	79.8	63.1	48.2	36.4	27.8	57.1	122.2	21.5	69.2	51.3	37.6	27.7	22.1	49.6	57.2
JRAN (BS=3) (ours)	80.8	64.4	49.6	38.3	28.4	58.2	128.0	21.8	69.9	53.0	37.9	27.9	24.8	52.6	57.9
JRAN (BS=4) (ours)	80.9	64.6	49.7	38.5	**28.5**	**58.4**	128.2	22.0	71.0	53.1	38.1	28.1	24.8	52.7	58.0
JRAN (BS=5) (ours)	81.0	**64.7**	**49.8**	**38.6**	28.3	58.3	**128.4**	**22.1**	**71.2**	**53.3**	**38.3**	**28.3**	**25.0**	**52.9**	**58.2**

Table 2: Performance of JRAN and related state-of-the-arts on two datesets. “BS" denotes the Beam Search strategy.

Methods	B-1	B-2	B-3	B-4	M	R	C	S
ARL (Wang et al., 2020)	75.9	60.3	46.5	35.8	27.8	56.4	111.3	-
KMSL (Li and Jiang, 2020)	79.2	63.2	48.3	36.3	27.6	56.8	120.2	21.4
JRAN (ours)	**81.0**	**64.7**	**49.8**	**38.6**	**28.3**	**58.3**	**128.4**	**22.1**

Table 3: Performance comparision of our JRAN with the similar relationship based methods on MSCOCO.

Methods	B-1	B-2	B-3	B-4	M	R	C	S
Transformer (Sharma et al., 2018)	80.2	64.8	**50.5**	38.6	28.8	58.5	128.3	**22.6**
VORN (Herdade et al., 2019)	80.5	-	-	38.6	28.7	58.4	128.3	**22.6**
LBPF (Qin et al., 2019)	80.5	-	-	38.3	28.5	58.4	127.6	22.0
Sim-Trans ♯	79.4	64.5	49.1	38.5	28.0	58.1	125.5	21.7
JRAN-Trans (ours) (BS=3)	**81.2**	**65.3**	50.1	**39.2**	**28.9**	**58.7**	129.4	**22.6**
JRAN-Trans (ours) (BS=4)	81.1	65.2	50.0	39.1	28.8	58.6	**129.6**	22.5
JRAN-Trans (ours) (BS=5)	80.9	65.0	49.8	39.0	28.5	58.5	129.3	22.2

Table 4: Performance comparisons with transformer based baseline model on MSCOCO dataset. ‘♯’ is the current mainstream transformer based baseline model.

upgraded model “JRAN-Trans" under different BS. Since the BLEU4 score of the original baselines “*Sim-Trans* " is 38.5, our upgraded transformer variation based model can effectively boost the score by 0.7. In addition, the CIDEr score is significantly increased from 125.5 to 129.6, which is clearly a meaningful improvement. Moreover, compared with the other current mainstream methods, our method still achieves very competitive performance across most metrics.

5.4 Model Accuracy and Efficiency

We further conduct an experimental computational cost analysis for comparing our updated model “JRAN-Trans" with some typical models (i.e., “SC (base) (Chen et al., 2017)", “Sim-Trans", and “X-Transformer" (Pan et al., 2020)). Figure 6 presents the computational cost in terms of the training time and parameters of the model.

As can be seen from Figure 6, the baseline model “SC (base)" has the less parameters and training

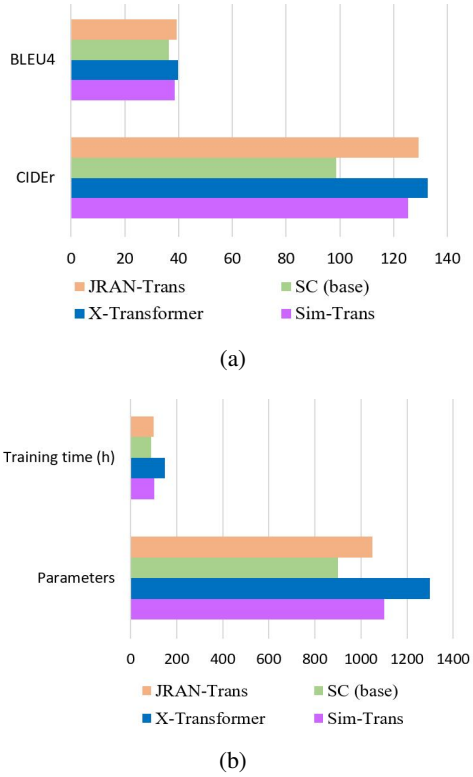


Figure 6: Illustration of the model computational cost (a) and model accuracy (b).

time, but the model accuracy is the lowest. One main reason is that it ignores the potential relationship between features in the image. In addition, the evaluation metric scores of the transformer based model “X-Transformer" are relatively high, while the parameter and training time are also relatively high.

Compared with the “X-Transformer" model, obviously, our “JRAN-Trans" does not significantly increase the computational cost of the model while achieving promising accuracy. This clearly indicates that our method displays a trade-off between

model accuracy and computational cost.

5.5 Qualitative Results

a. Comparison of Generating Sentences with Different Approaches: Establishing the interaction between visual and semantic information by learning the relationship among image features, our JRAN can generate comprehensive sentences more consistent with the image theme scene. Figure 7 shows some examples of sentence description generated by different baseline methods, namely Baseline (Li et al., 2018a), *Sim-Trans* and our JRAN-Trans.

Figure 7: Examples of generating sentences. Blue is region objects, purple is the background or environment information of the object in the image, the red is the corresponding visual relationship, and the underline indicates the corresponding visual-semantic relationship.

From these exemplar results, it is clearly see that the three methods can generate somewhat relevant and logically correct sentences, while our JRAN based method "JRAN-Trans" generates more consistent sentences with image theme scene. It effectively improves the quality of generated text by enriching visual-semantic relationships. For instance, for image (a), compared to the relationship words/phrases "playing" and "playing with" generated by methods "baselines" and "Sim-Trans" respectively. Our "JRAN-Trans" not only accurately generates the relationship phrase "staring at" between region objects "dogs", but also enrichs the relationship between region object "dog" and its background word "road", and generates an appropriate interactive word "on" between them, which significantly improves the model overall performance.

b. Effectiveness of Relationship Gate: Figure 8 visualizes the weight of relationship gate during generating different words. It can be seen that visual relationship and visual-semantic relationship contribute differently to the generation of different words. Specifically, the visual relationship has

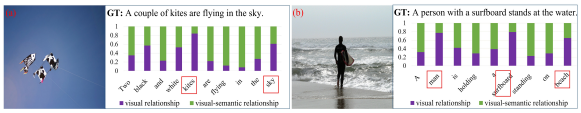


Figure 8: Visualization of relationship gate weight. The first column is original image, and the second column are the values of visual relationship weight (purple) and visual-semantic relationship weight (green).

a larger value when generating the visual words "kite" and "sky" (red boxes). Contrarily, when non-visual words (like "Two", "and", "flying" etc.) are generated, the visual-semantic relationship has greater value. This indicates that the two relationships complement each other and jointly boost the performance of image-text generation.

6 Conclusion

In the paper, a simple and effective model that makes full use of the complementary region and semantic features in the image, Joint Relationship Attention Network (JRAN) is proposed. It explores the relationship among the features to enrich the relationship-level representation for finally boosting image-text generation. To verify our claim, we propose a new joint relationship learning network, which is able to learn two kinds of feature relationships. Considering the different contributions of these two relationships during generating words, we further devise a relationship gate to finally obtain a feature representation containing different-level relationship information. Importantly, our model has made remarkable progress in deeply exploring the relationship between features for image-text generation. Extensive experiments demonstrate that the effectiveness of our proposed model on larger-scale and smaller-scale datasets. More remarkably, we obtain new state-of-the-art performances on popular Flickr30k dataset.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under grant 62176062.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *Proceedings European Conference on Computer Vision*, pages 382–398.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, page 6077–6086.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of Meeting of the Association for Computational Linguistics*, pages 65–72.

Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. 2017. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 5659–5667.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, page 10578–10587.

Yan Duan, Marcin Andrychowicz, Bradly Stadie, Jonathan Ho, Jonas Schneider, Ilya Sutskever, Pieter Abbeel, and Wojciech Zaremba. 2017. One-shot imitation learning. In *arXiv:1703.07326*.

Junlong Gao, Shiqi Wang, Shanshe Wang, Siwei Ma, and Wen Gao. 2019. Self-critical n-step training for image captioning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, page 6300–6308.

Xinwei He, Yang Yang, Baoguang Shi, and Xiang Bai. 2019. Vd-san: visual-densely semantic attention network for image caption generation. *Neurocomputing*, 328:48–55.

Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2019. Image captioning: Transforming objects into words. In *Proceedings of Advances in Neural Information Processing Systems*, page 11135–11145.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *Proceedings of International Conference on Learning Representations*.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, and et al. Ivan Krasin. 2018. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982*.

Linghui Li, Sheng Tang, Yongdong Zhang, Lixi Deng, and Qi Tian. 2018a. Gla: Global-local attention for image description. *IEEE Transactions on Multimedia*, 20(3):726–737.

Ruoyu Li, Sheng Wang, Feiyun Zhu, and Junzhou Huang. 2018b. Adaptive graph convolutional neural networks. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 3546–3553.

Xiangyang Li and Shuqiang Jiang. 2020. Know more say less: Image captioning based on scene graphs. *IEEE Transactions on Multimedia*, 20(8):2117 – 2130.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of Meeting of the Association for Computational Linguistics*, pages 74–81.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: common objects in context. In *Proceedings of European Conference on Computer Vision*, pages 740–755.

Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. 2020. X-linear attention networks for image captioning. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10968–10977.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of Meeting of the Association for Computational Linguistics*, pages 311–318.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of International Conference on Computer Vision*, pages 2641–2649.

Yu Qin, Jiajun Du, and Hongtao Zhang, Yonghuaand Lu. 2019. Look back and predict forward in image captioning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, page 8367–8375.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of 56th Annual Meeting of the Association-for-Computational-Linguistics*, pages 2556–2565.

Jingkuan Song, Pengpeng Zeng, Lianli Gao, and Heng Tao Shen. 2018. From pixels to objects: Cubic visual attention for visual question answering. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 906–912.

Sahar Tahvili, Leo Hatvani, Enislay Ramentol, Rita Pimentel, Wasif Afzal, and Francisco Herrera. 2020. A novel methodology to classify test cases using natural language processing and imbalanced learning. *Engineering Applications of Artificial Intelligence*, 95:103878.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *arXiv preprint arXiv:1706.03762*.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575.

Cheng Wang, Haojin Yang, Christian Bartz, and Christoph Meinel. 2016. Image captioning with deep bidirectional lstms. In *Proceedings of the 2016 ACM Multimedia Conference*, pages 988–997.

Junbo Wang, Wei Wang, Liang Wang, Zhiyong Wang, David Dagan Feng, and Tieniu Tan. 2020. Learning visual relationship and context-aware attention for image captioning. *Pattern Recognition*, 98:107075.

Weixuan Wang, Zhihong Chen, and Haifeng Hu. 2019. Hierarchical attention network for image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8957–8964.

Jie Wu, Tianshui Chen, Hefeng Wu, Zhi Yang, Guangchun Luo, and Liang Lin. 2021a. Fine-grained image captioning with global-local discriminative objective. *IEEE Transactions on Multimedia*, 23:2413–2427.

Lingxiang Wu, Min Xu, Lei Sang, Ting Yao, and Tao Mei. 2021b. Noise augmented double-stream graph convolutional networks for image captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(8):3118–3127.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning*, pages 2048–2057.

Liang Yang, Haifeng Hu, Songlong Xing, and Xinlong Lu. 2020. Constrained lstm and residual attention for image captioning. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 16(3):1–18.

Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2019. Hierarchy parsing for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, page 2621–2629.

Senmao Ye, Junwei Han, and Nian Liu. 2018. Attentive linear transformation for image captioning. *IEEE Transactions on Image Processing*, 27(11):5514–5524.

Yu Zhang, Xinyu Shi, Siya Mi, and Xu Yang. 2021. Image captioning with transformer and knowledge graph. *Pattern Recognition Letters*, 143:43–49.