ER-SAN: Enhanced-Adaptive Relation Self-Attention Network for Image Captioning

Jingyu Li[1] , **Zhendong Mao**[1*] , **Shancheng Fang**[1] and **Hao Li**[2]
[1]University of Science and Technology of China, Hefei, China
[2]Huazhong University of Science and Technology, Wuhan, China
jingyuli@mail.ustc.edu.cn, {zdmao, fangsc}@ustc.edu.cn, hao_li_cn@hust.edu.cn

Abstract

Image captioning (IC), bringing vision to language, has drawn extensive attention. Precisely describing visual relations between image objects is a key challenge in IC. We argue that the visual relations, that is geometric positions (i.e., distance and size) and semantic interactions (i.e., actions and possessives), indicate the mutual correlations between objects. Existing Transformer-based methods typically resort to geometric positions to enhance the representation of visual relations, yet only using the shallow geometric is unable to precisely cover the complex and actional correlations. In this paper, we propose to enhance the correlations between objects from a comprehensive view that jointly considers explicit semantic and geometric relations, generating plausible captions with accurate relationship predictions. Specifically, we propose a novel Enhanced-Adaptive Relation Self-Attention Network (ER-SAN). We design the direction-sensitive semantic-enhanced attention, which considers content objects to semantic relations and semantic relations to content objects attention to learn explicit semantic-aware relations. Further, we devise an adaptive re-weight relation module that determines how much semantic and geometric attention should be activated to each relation feature. Extensive experiments on MS-COCO dataset demonstrate the effectiveness of our ER-SAN, with improvements of CIDEr from 128.6% to 135.3%, achieving state-of-the-art performance. Codes will be released https://github.com/CrossmodalGroup/ER-SAN.

1 Introduction

Image captioning, aiming to automatically generate descriptions for a given image, is a crucial multi-modal task since it brings vision to language. The generated caption is expected to not only recognize the interested objects, but also describe the relationships between image objects. For image captioning, a key challenge is how to precisely and effectively model the relationships between recognized image objects, which is essential to improving generation quality.

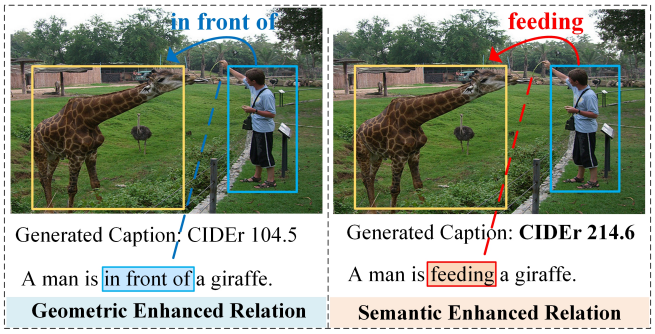


Figure 1: Comparison of generated relations enhanced by geometric vs. semantic. Existing methods (left) with geometrical enhancement typically generate relation "in front of" for the given image, yet the shallow geometry is insufficient to cover the complex and actional correlations. Our method (right) further exploits the semantic enhancement to generate key semantic relation "feeding", i.e., capturing more precise visual semantic associations in generated captions.

The Transformer-based self-attention model has achieved superior performances for image captioning, yet the self-attention only implicitly calculates the similarity of a single region w.r.t. all other regions. Recently, some researchers focus on variants or modifications of the self-attention operator via using geometric information to enhance region-level features [Huang *et al.*, 2019; Pan *et al.*, 2020]. Since the geometric features, i.e., relative distance and relative size, contain the explicit position relations between objects, approaches leveraging the geometric features in self-attention achieve state-of-the-art performance [Herdade *et al.*, 2019; Guo *et al.*, 2020; Song *et al.*, 2021]. These facts clearly show that a promising direction is to design the visual encoder that expects to accurately model the relations between image objects.

We argue that the visual relations, referring to geometric positions and semantic interactions (i.e., actions and possessives) [Yao *et al.*, 2018], indicate the mutual correlations between objects in region-level representations. However, existing works only resort to geometric positions to enhance the representation of visual relations, in which the shallow geometric is unable to precisely cover the complex and actional

*Zhendong Mao is the corresponding author

correlations. As a result, they suffer from the limitation that it is difficult for the caption model to generate plausible captions with accurate relation predictions. Compared with the obvious geometric relations, the semantic relations between image objects are often abstract, typically defined by the precise visual semantic associations. For example, as illustrated in Fig. 1 (a), resorting to the geometric-enhanced region features representation, the caption model generates the obvious geometric relation "in front of" but lacks pivotal semantic relation. In contrast, using the explicit semantic guidance in Fig. 1 (b), the caption model will generate a more accurate relation prediction "feeding", where the caption enjoys better CIDEr performance (the higher, the better).

To address the above issue, we propose a novel Enhanced-Adaptive Relation Self-Attention Network (ER-SAN) for image captioning to model the relationships on both explicit semantic and geometric levels, which integrates the connections into the image encoder to enrich relation-aware region-level representations. ER-SAN adaptively enhances the visual correlations in the given image when there are explicit relationships between two objects. Specifically, our method contains three modules 1) Feature Extraction: we first extract visual semantic associations, salient regions and bounding boxes to construct the semantic and geometric graph. 2) Enhanced-Adaptive Relation Attention: we devise the enhanced attention with relation-aware bias including semantic-enhanced attention, geometric-enhanced attention and content weight. To learn explicit semantic-enhanced relations, we design a novel direction-sensitive semantic-enhanced attention, which considers content objects to semantic relations and semantic relations to content objects attention to leverage directional semantic relation information. To learn geometric-enhanced relations, we conduct dynamical geometric-enhanced attention between content and geometric features. Moreover, ER-SAN would adaptively enhance different levels of relation features for different geometric and semantic dependencies. 3) Language Prediction: after encoding, we leverage the transformer decoder to generate plausible captions.

Our contributions are summarized as follows:

- We propose a novel Enhanced-Adaptive Relation Self-Attention Network, which is the first time to, for the best of our knowledge, leverage semantic and geometric relation to jointly cover complex visual relations between objects in the transformer-based model.
- We devise an adaptively enhanced semantic and geometric relation method when there are explicit relationships between two objects and leverages the Transformer's ability to learn the new relations.
- We conduct extensive experiments on MS-COCO dataset, and achieve a new state-of-the-art performance for image captioning, i.e., 135.3% CIDEr score and 23.8% SPICE score on Karpathy test set.

2 Related Work

2.1 Transformer for Image Captioning

Typically, image captioning models use CNN to extract visual features from an image and RNN to generate the corresponding output descriptions. The bottom-up and top-down attention proposed by Anderson et al. [Anderson *et al.*, 2018] uses Faster R-CNN [Ren *et al.*, 2015] proposing image regions as an additional bottom-up path coupled with the grid-level features to upgrade the visual features. Self-attention was first proposed by Vaswani et al. [Vaswani *et al.*, 2017] for machine translation task. Plenty of works proposed variants or modifications of the self-attention mechanism, which have successfully achieved superior performances in image captioning [Luo *et al.*, 2021; Zhang *et al.*, 2021]. For example, Huang et al. [Huang *et al.*, 2019] introduced "Attention on Attention" in which the context gate guides the final attended information. On a related line, Cornia et al. [Cornia *et al.*, 2020] proposed the augmented self-attention with a set of memory vectors of each encoder layer.

2.2 Explicit Relation for Image Captioning

Some prior studies consider using Graph Convolutional Nets (GCN) [Kipf and Welling, 2016] to encode objects and their relationships. Yao et al. [Yao *et al.*, 2018] first attempt to use semantic relationships graph, which is obtained by a pre-trained semantic classifier, and spatial relationships graph, which is captured between bounding boxes of object pairs. Following this work, Yang et al. [Yang *et al.*, 2019] proposed to integrate semantic priors learned from the text in image encoding by exploiting a graph-based representation of both images and sentences. This retrieved graph is then supplied to a GCN for the encoder and reuse at the decoding stage.

Such methods have two technical limitations: (1) encoding semantic and geometric relations independently and concatenating them via a simple ensemble; (2) making use of solely explicit correlations while ignoring implicit correlations. In contrast, our approach adaptively fuses semantic and geometric clues by the attention mechanism at each Transformer layer, and considers not only explicit correlations but also implicit correlations (via fully connected self-attention).

3 Methodology

Given an image I, the task of image captioning is to automatically generate a sentence S based on the visual entities in the image. Our goal is to take full advantage of the explicit relation as guidance to generate more accurate captions. The overall framework of our proposed ER-SAN is depicted in Fig. 2. The proposed method can be divided into three sub-modules: (a) Feature Extraction Module (Sec. 3.1), (b) Enhanced-Adaptive Relation Attention (Sec. 3.2) and (c) Language Prediction Module (Sec. 3.3). Concretely, we first extract visual, semantic and geometric features to jointly model image representation. Next, we explicitly enhance region-level relations. Last, our model generates plausible captions with accurate relationship predictions.

3.1 Feature Extraction Module

We use object features, semantic graph and geometric graph features to jointly represent the complex relations between objects. Object feature has rich object detail information, which is essential for image understanding. Semantic graph represents the linguistic actions between objects. Geometric

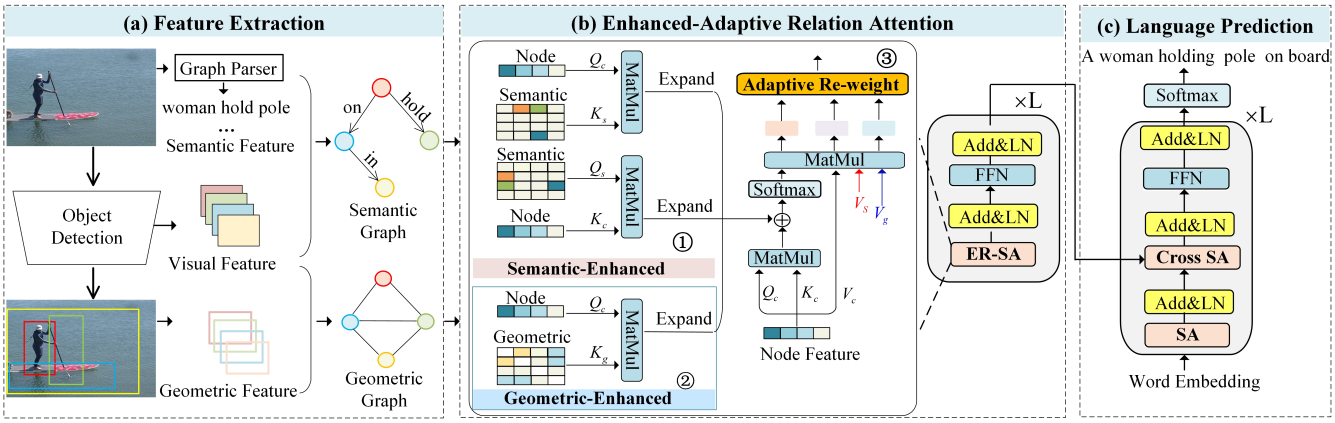


Figure 2: An overview of our ER-SAN, which consists of three modules: (a) Feature Extraction: the graph parser is employed to extract visual semantic associations to construct the semantic graph, and Faster-RCNN detects the salient regions and bounding boxes to construct the geometric graph. (b) Enhanced-Adaptive Relation Attention: 1) the direction-sensitive semantic-enhanced considering content objects to semantic relations and semantic relations to content objects attention to jointly represent complete triples information; 2) the geometric-enhanced dynamically computes attention between content and geometric features; 3) the adaptive relation re-weight adaptively enhances different relation features. (c) Language Prediction: leveraging the transformer decoder predicts plausible captions with accurate relations.

graph reflecting the spatial patterns of individual objects can complement the visual information.

Contextual Node Feature

We firstly utilize Faster R-CNN to encode the region-level RoI features into n region features $o^v=[o_1^v, o_2^v, \ldots, o_n^v]$. Also, we obtain the object category labels c, and then we adopt the learnable object embedding layer to map the object category labels c into feature embedding vectors o^c. To obtain the contextual representation of object node feature, we concatenate visual feature o^v and embedding vectors o^c as:

$$H_{o_i} = \phi_o([o^v; o^c]) + o^v, \tag{1}$$

where $H_{o_i} \in \mathbb{R}^{N\times d_k}$ is the next input node region feature representation passing in the transformer encoder layer, ϕ_o is feed-forward networks using ReLU activation, [;] means concatenation operation.

Semantic Graph Construction

To represent the precise visual semantic associations, we use the graph parser to extract the semantic relation triples (e.g. $<subject\text{-}predicate\text{-}object>$) following [Shi *et al.*, 2020], which constructs caption-guided visual relationship graphs that introduce beneficial inductive bias using weakly supervised multi-instance learning. We utilize the semantic relation triples to calculate the semantic relation matrix between subject region i and object region j, its dimensions are $\mathbb{R}^{N\times N}$. Here, the semantic relation matrix is not symmetric. For example, as a triple $<woman\text{-}holding\text{-}umbrella>$ is a proper semantic relation $holding$, while there is no semantic relation $holding$ from $umbrella$ to $woman$. Also, if two objects are not semantic related, we set the element to be a special value, i.e., 0.

We adopt the learnable semantic embedding layer to map the semantic relation matrix into high-dimensional semantic vectors $S \in \mathbb{R}^{N\times N\times d_s}$ whose each element S_{ij} represents a d_s-dimensional semantic embedding between region i and j. This explicit semantic relation embedding will be as prior knowledge to guide model encode more accurate relation.

Geometric Graph Construction

Different from the semantic graph, the geometric graph is undirected. We first calculate relative geometric features between two boxes i and j as g_{ij} which is a 4-dimensional vector containing relative distance and size [Guo *et al.*, 2020]

$$(\log(\frac{|x_i - x_j|}{w_i}), \log(\frac{|y_i - y_j|}{h_i}), \log(\frac{w_i}{w_j}), \log(\frac{h_i}{h_j})), \tag{2}$$

where $g_{ij} \in \mathbb{R}^{N\times N\times 4}$, x, y, w and h denote the box's center coordinates and its width and height, respectively.

To get the geometric embedding, we use the PE_{pos} functions [Vaswani *et al.*, 2017] to calculate high-dimensional representations $G \in \mathbb{R}^{N\times N\times d_g}$. Also, to get sparse geometric graph, if the relative distance $dis(b_i, b_j)$ and Intersection over Union $IoU(b_i, b_j)$ of two objects are within a given threshold, the geometric relation between them is assigned.

3.2 Enhanced-Adaptive Relation Attention

In this section, our goal is to model relations between objects leveraging the semantic and geometric embedding information of graphs into the Transformer model. Therefore, we propose a novel Enhanced-Adaptive Relation Attention which considering semantic-enhanced and geometric-enhanced mechanisms adaptively encode arbitrary relations between pairwise objects, and these representations are jointly computed over all inputs using self-attention as shown in Fig. 2 (b).

Enhanced Relation Attention

To achieve the proposed enhanced relation attention, we extend the vanilla self-attention with relation-aware prior knowledge dividing into three parts: semantic-enhanced attention, geometric-enhanced attention and content weight.

Different from existing methods that ignore the semantic relations contained in images, this is the first time to incorporate explicit visual semantic associations into Transformer. Specifically, the enhanced attention score including the effect of content objects H, semantic relation embedding S and geometric relation embedding G computed as:

$$Q_c = HW_{q,c}, \quad K_c = HW_{k,c},$$
$$\tilde{A} = \underbrace{\boldsymbol{Q}_c K_c^\top}_{\text{Content Weight}} + \underbrace{\phi(\mathrm{Q}_c, \mathrm{K}_c, \mathrm{S})}_{\text{Semantic-Enhanced}} + \underbrace{\varphi(Q_c, \mathrm{G})}_{\text{Geometric-Enhanced}}, \quad (3)$$

where attention matrix $\tilde{A} \in \mathbb{R}^{N\times N}$, the scaling factor $\sqrt{d}$ is omitted for simplicity. ϕ is the semantic-enhanced attention function. φ is the geometric-enhanced attention function. Q_c and $K_c \in R^{N\times d}$ are the projected content objects vectors generated using projection matrices $W_{q,c}$ and $W_{k,c} \in R^{d_k\times d}$, respectively. In Eq.(3), both implicit relations (content weight) and explicit relations guidance (semantic and geometric) are considered. In the following, we will introduce these two functions $\phi(\mathrm{Q}_c, \mathrm{K}_c, \mathrm{S})$ and $\varphi(Q_c, \mathrm{G})$ in detail.

Direction-Sensitive Semantic-Enhanced Attention. Different from existing methods, we propose a novel semantic-enhanced attention, which uses semantic embedding features to enhance attention score when there is a specific relationship between two objects. As discussed in Sec. 3.1, relationships in the semantic graph are directional. For solving the problem: *how to use the directional relation jointly inferring complete triples information?* We propose a novel direction-sensitive semantic-enhanced attention. Specifically, we consider content objects to semantic relations (c2s) and semantic relations to content objects (s2c) attention. For example, for a directional relation triple $<$*woman-holding-umbrella*$>$, which can be decomposed into $<$*woman-holding*$>$ and $<$*holding-umbrella*$>$. To this end, we conduct the directed semantic relation guidance:

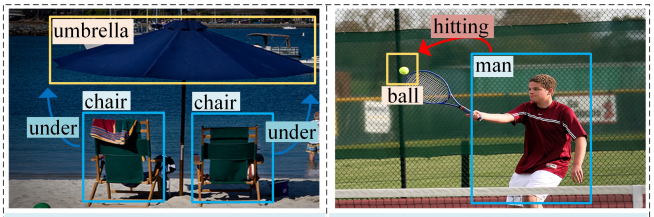
$$Q_s = SW_{q,s}, \quad K_s = SW_{k,s},$$
$$\phi(\mathrm{Q}_c, \mathrm{K}_c, \mathrm{S}) = \underbrace{Q_c K_s^\top}_{\text{c2s}} + \underbrace{Q_s K_c^\top}_{\text{s2c}}, \quad (4)$$

where Q_c and K_s are the projected content vectors and semantic embedding, subscripts c and s represent content and semantic, respectively. $W_{q,s}$ and $W_{k,s} \in R^{d_s\times d}$ are the projection matrices.

Note that to match the dimension of semantic attention with the content weight, we perform the following dimension transformation. Taking content objects to semantic relations an example, for multi-heads h, the dimension of content Q_c is $R^{h\times N\times d}$. We first permutate Q_c as $R^{N\times h\times d}$ and semantic $K_s^\top$ as $R^{N\times d\times N}$. Then, we compute dot product which means the relation vectors sharing across heads.

Geometric-Enhanced Attention. Different from semantic relation, geometric relation matrix is symmetric. Inspired by [Guo *et al.*, 2020], we conduct a dynamical geometric-enhanced attention between content Q_c and geometric relation K_g as:

$$K_g = GW_{k,g}, \quad \varphi(\mathrm{Q}_c, \mathrm{G}) = Q_c K_g^\top. \quad (5)$$



(a) Geometric Relation Concerned (b) Semantic Relation Concerned

Figure 3: Illustration of the dependence of caption generation on semantic and geometric information at different scenes. For the image (a) with obvious geometrical dependent, e.g., "under" while the image (b) with semantical dependent, e.g., "hitting", our method devises the adaptive mechanism to re-weight corresponding dependencies, which provides accurate relation generation guidance.

Adaptive Re-weight Relation

We observe inconsistent dependence of caption generation on semantic and geometric information at different scenes, which indicates that the attention should be taken into account to determine when the semantic and geometric features need to be activated. For example, in Fig. 3 (a), there is a clear geometric relationship, e.g. "chair under umbrella". It is appropriate to pay more attention to geometric relation, which is called geometric relation concerned. In contrast, in Fig. 3 (b), there is no obvious geometric relation to representing the man's action "hitting". It is necessary to pay more attention to semantic enhanced relation to guide caption model generate more accurate relation predictions, which is called semantic relation concerned.

To this end, we propose an adaptive learning re-weight module in the attention aggregation operation to get different levels of enhanced semantic and geometric relation information. We introduce the gated fusion module and conduct sigmoid function $\sigma(\cdot)$ as:

$$V_c = HW_{v,c}, V_s = SW_{v,s}, V_g = GW_{v,g},$$
$$\beta = \sigma\left(W_\beta\left[V_c, V_s, V_g\right]\right), \quad (6)$$

where $\beta = [\beta^s, \beta^g] \in \mathbb{R}^{N\times 2}$, W_β are learnable parameters. The gating value indicates which feature information is more important for the current state. The output of enhanced-adaptive relation attention is computed as:

$$Z = \mathrm{Softmax}(\widetilde{A})\left(V_c + \beta^s \cdot V_s + \beta^g \cdot V_g\right). \quad (7)$$

Then, we can accumulate the final explicit exploration representation followed by an FFN operation as Fig. 2 (b).

3.3 Language Prediction Module

After encoding, we use the Transformer decoder for caption generation. The inputs of language prediction module are features from the last encoder layer, which contain aligned relations between region features and words, while the outputs are descriptive sentences shown as Fig. 2 (c). The enhanced-adaptive relation attention is not used in language prediction module because the decoder is autoregressive and inputs variable-length sentences. After decoding layer, it outputs $H^{\mathcal{D}}$. Then, the confidence of the word distribution is:

$$\mathbf{P}(\mathcal{S}) = \mathrm{Softmax}\left(\mathbf{W}\mathcal{H}^{\mathcal{D}} + b\right). \quad (8)$$

Model	Cross Entropy						RL					
	B@1	B@4	M	R	C	S	B@1	B@4	M	R	C	S
	Single Model											
GCN-LSTM [Yao *et al.*, 2018]	77.3	36.8	27.9	57.0	116.3	20.9	80.5	38.2	28.5	58.3	127.6	22.0
MT-I [Shi *et al.*, 2020]	78.1	38.4	28.2	58.0	119.0	21.1	80.8	38.9	28.8	58.7	129.6	22.3
NG-SAN [Guo *et al.*, 2020]	-	-	-	-	-	-	-	39.9	29.3	59.2	132.1	23.3
X-Linear [Pan *et al.*, 2020]	-	-	-	-	-	-	80.9	39.7	29.5	59.1	132.8	23.4
TCIC [Fan *et al.*, 2021]	78.1	38.3	28.5	58.0	121.0	21.6	80.9	39.7	29.2	58.6	132.9	22.4
DRT [Song *et al.*, 2021]	-	-	-	-	-	-	81.7	40.4	29.5	59.3	133.2	23.3
Transformer (Base)	75.8	34.2	27.7	56.1	113.4	21.0	80.6	38.4	28.6	58.4	128.6	22.6
ER-SAN[Ours]	**78.2**	**38.8**	**29.2**	**58.5**	**122.9**	**22.2**	**82.1**	**41.7**	**30.1**	**60.3**	**135.3**	**23.8**
	Ensemble Model											
GCN-LSTM$^{\Sigma}$	77.4	37.1	28.1	57.2	117.1	21.1	80.9	38.3	28.6	58.5	128.7	22.1
X-Linear$^{\Sigma}$	77.8	37.7	29.0	58.0	122.1	21.9	81.7	40.7	29.9	59.7	135.3	23.8
TCIC$^{\Sigma}$	78.8	39.1	29.1	58.5	123.9	22.2	81.8	40.8	29.5	59.2	135.3	22.5
ER-SAN (Ours)	**79.0**	**39.6**	**29.6**	**59.0**	**125.1**	**22.6**	**83.0**	**43.1**	**30.5**	**60.9**	**137.6**	**24.2**

Table 1: Comparisons with state-of-the-art single and ensemble model on MS-COCO Karpathy test split. B@1, B@4, R, M, C and S are denoted for BLEU-1, BLEU-4, ROUGE, METEOR, CIDEr and SPICE, respectively. Σ represents ensemble model. The bests are in **bold**.

4 Experiments

4.1 Dataset and Implementation Details

Dataset. To validate our proposed framework, we conduct extensive experiments on the MS-COCO [Lin *et al.*, 2014] which is the most commonly used dataset for image captioning. It contains 123,287 images, each of them annotated at least with five different captions. According to the Karpathy splits [Karpathy and Fei-Fei, 2015], we split 5,000 images are used for validation, 5,000 images for testing, and 113,287 images for training.

Evaluation Metrics. We use five standard automatic evaluation metrics to evaluate the quality of image captioning, including BLEU [Papineni *et al.*, 2002], METEOR [Banerjee and Lavie, 2005], ROUGR [ROUGE, 2004], CIDEr [Vedantam *et al.*, 2015], and SPICE [Anderson *et al.*, 2016].

Implementation Details. Following the Transformer-based model [Vaswani *et al.*, 2017] and [Guo *et al.*, 2020], our encoder and decoder have same layers L, the number of heads is 8, the hidden dimension is 512, the inner dimension of feed-forward module is 2048 and $dropout = 0.1$. We use Adam optimizer with a mini-batch size of 10 to train our model. For cross entropy training, we increase the learning rate linearly to $3e^{-4}$ with warm-up for 3 epochs, and then decay by rate 0.5 every 3 epochs. We first train the model for 18 epochs with the cross-entropy loss and then further optimize with CIDEr reward for additional 40 epochs with a fixed learning rate value of $5e^{-6}$. The size of beam search is 2 to generate captions during testing. If not specifically specified, we set the baseline transformer model layers $L = 4$.

4.2 Comparison with state-of-the-arts

Model Comparison. For fair comparison, we use the RCNN-based region features and compare our models with the following state-of-the-arts: GCN-LSTM [Yao *et al.*, 2018], MT-I [Shi *et al.*, 2020], X-Transformer [Pan *et al.*, 2020], NG-SAN [Guo *et al.*, 2020], TCIC [Fan *et al.*, 2021] and DRT [Song *et al.*, 2021].

Single Model. In Table 1, we list the two results (Cross Entropy and RL) performance of our method in comparison with state-of-the-art methods on the MSCOCO Karpathy split. As it can be observed, our model outperforms the baseline model Transformer ($L = 4$) significantly, with CIDEr scores being improved from 113.4 to 122.9 (cross-entropy) and 128.6 to 135.3 (RL) in the CIDEr score. Our method surpasses all other approaches. Focusing on the BLEU4 metric, ER-SAN absolutely outperforms DRT models by 3.2%, which indicates ER-SAN generates more accurate phrases.

Ensemble Model. We build an ensemble of our models by averaging the output probability distributions of several independently trained instances of the model, and these instances are trained from different random seeds. Noticeably, our ensemble achieves the best performance on all metrics.

4.3 Ablation Study

Analysis on ER-SAN. To investigate the relative contributions of different components in ER-SAN, we perform the extensive ablation studies in Table 2, which is trained with cross-entropy for 18 epochs. T represents the Transformer ($L = 4$). GE, SE, AR stand for Geometric-Enhanced Attention, Direction-Sensitive Semantic-Enhanced Attention and Adaptive Re-weight in Sec. 3.3, respectively. SE (c2s) and SE (s2c) denote the content objects to semantic relations and the semantic relations to content objects in Eq.(4), respectively. The proposed model amounts to a total improvement of 9.5 points CIDEr from 113.4 to 122.9 with respect to the base Transformer. In addition, we observe a significant increase in performance from 113.4 to 120.0 CIDEr, when only adding SE module indicating the effectiveness of the semantic enhanced relation. Finally, combining GE, SE, AR units

Model	B@1	M	R	C	S
T (base)	75.8	27.7	56.1	113.4	21.0
T+GE	76.1	28.2	56.7	116.1	21.2
T+SE (c2s)	76.8	28.7	57.5	118.3	21.7
T+SE (s2c)	77.3	28.8	57.7	119.2	21.9
T+SE	77.6	28.9	58.0	120.0	21.9
T+GE+SE	77.8	29.1	58.3	121.4	22.1
T+GE+SE+AR	**78.2**	**29.2**	**58.5**	**122.9**	**22.2**

Table 2: Ablation study for the proposed ER-SAN model.

β^s	β^g	B@1	B@4	M	R	C	S
0	0	76.2	36.6	28.3	57.1	117.1	21.3
0	1	76.2	36.5	28.3	56.9	116.8	21.3
1	0	77.2	38.1	28.9	58.0	119.9	21.6
1	1	77.8	38.4	29.1	58.3	121.4	22.1
AR		**78.2**	**38.8**	**29.2**	**58.5**	**122.9**	**22.2**

Table 3: Ablation study for the adaptive re-weight relation.

altogether, we get the highest performance on all metrics.

Analysis on Adaptive Re-weight Relation. In order to further analyze the effectiveness and necessity of AR moudle. In Table 3, we fix the $\tilde{A}$ in Eq.(3) to change the β^s and β^g in Eq.(7). $\beta^s = 0$ and $\beta^g = 0$ mean without considering semantic features V_s and geometric features V_g, respectively. From the first and fourth rows of the table, it is necessary to the aggregate attention operation considering the semantic V_s and geometric V_g features. Note that our adaptive learning enhanced relation module better than all fixed β, it indicates that inconsistent dependence of caption generation on semantic and geometric information.

Analysis on Relation Improvement. For better understanding the improved relation performance, we list the breakdown of SPICE metric, which can measure how well caption models recover objects, attributes and relations in Table 4. We performed a more detailed comparison of the vanilla Transformer (T), Transformer with geometric-enhanced (T+GE), Transformer with semantic-enhanced (T+SE) against their combination (T+GE+SE). Compared to T+GE, T+SE has significantly higher relation scores, which indicates semantic-enhanced will precisely cover more complex and actional correlations. With our expectation that adding the geometric and semantic enhanced attention (T+GE+SE) would jointly help the model predict correct relations between objects.

Model	SPICE	Object	Relation	Attribute
T	21.0	37.8	5.9	11.2
T+GE	21.2	37.9	6.3	11.4
T+SE	21.9	38.7	7.1	11.6
T+GE+SE	**22.1**	**39.1**	**7.4**	**11.9**

Table 4: Breakdown of SPICE metrics.

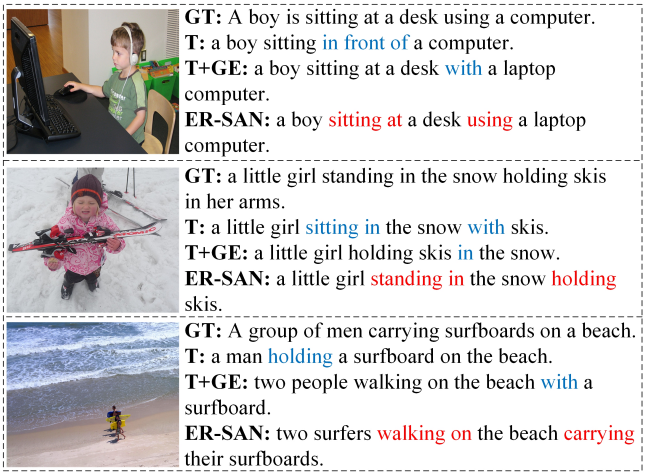


Figure 4: Examples of the ground-truths captions, as well as generated captions by the Transformer-based, Transformer with geometric enhanced and ER-SAN model corresponding. The blue and red words represent the generated inaccurate relations and precise semantic relations, respectively.

4.4 Qualitative Results and Visualization

Fig. 4 illustrates several example image captions generated by the base Transformer, Transformer with geometric relation and ER-SAN. The wrong phrases and accurate relation are highlighted in blue and red separately. We can observe that our model ER-SAN can generate more accurate semantic relation, e.g. "using", "holding" and "carrying". Note that the ER-SAN with a comprehensive understanding of the visual relations can generate "boy using computer".

5 Conclusion and Future Work

In this paper, we propose a novel enhanced relation framework, which enhances the correlations between objects from a comprehensive view that jointly considers explicit semantic and geometric relations, generating plausible captions with accurate relationship predictions. Different from existing works, we design direction-sensitive semantic-enhanced attention to learn explicit semantic-aware relations. Besides, the proposed gated fusion module can adaptively re-weight semantic and geometric relation features for caption generation. Extensive experiments demonstrate the superiority of our ER-SAN. For further works, we tend to exploit this general relation-enhanced attention to other multi-modal tasks, such as visual question answering, visual grounding.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (2020YFB1406603), the National Natural Science Foundation of China under Grants 62102384, China Postdoctoral Science Foundation (No. 2021M693092) and the Fundamental Research Funds for Central Universities under Grants WK3480000010 and WK3480000008.

References

[Anderson *et al.*, 2016] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer, 2016.

[Anderson *et al.*, 2018] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[Banerjee and Lavie, 2005] Satanjeev Banerjee and Alon Lavie. 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*, pages 65–72, 2005.

[Cornia *et al.*, 2020] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10578–10587, 2020.

[Fan *et al.*, 2021] Zhihao Fan, Zhongyu Wei, Siyuan Wang, Ruize Wang, Zejun Li, Haijun Shan, and Xuanjing Huang. Tcic: Theme concepts learning cross language and vision for image captioning. *arXiv preprint arXiv:2106.10936*, 2021.

[Guo *et al.*, 2020] Longteng Guo, Jing Liu, Xinxin Zhu, Peng Yao, Shichen Lu, and Hanqing Lu. Normalized and geometry-aware self-attention network for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10327–10336, 2020.

[Herdade *et al.*, 2019] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. *arXiv preprint arXiv:1906.05963*, 2019.

[Huang *et al.*, 2019] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4634–4643, 2019.

[Karpathy and Fei-Fei, 2015] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015.

[Kipf and Welling, 2016] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

[Luo *et al.*, 2021] Yunpeng Luo, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, Yongjian Wu, Feiyue Huang, Chia-Wen Lin, and Rongrong Ji. Dual-level collaborative transformer for image captioning. *arXiv preprint arXiv:2101.06462*, 2021.

[Pan *et al.*, 2020] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10971–10980, 2020.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015.

[ROUGE, 2004] Lin CY ROUGE. A package for automatic evaluation of summaries. In *Proceedings of Workshop on Text Summarization of ACL, Spain*, 2004.

[Shi *et al.*, 2020] Zhan Shi, Xu Zhou, Xipeng Qiu, and Xiaodan Zhu. Improving image captioning with better use of captions. *arXiv preprint arXiv:2006.11807*, 2020.

[Song *et al.*, 2021] Zeliang Song, Xiaofei Zhou, Linhua Dong, Jianlong Tan, and Li Guo. Direction relation transformer for image captioning. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5056–5064, 2021.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[Vedantam *et al.*, 2015] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[Yang *et al.*, 2019] Xu Yang, Hanwang Zhang, and Jianfei Cai. Learning to collocate neural modules for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[Yao *et al.*, 2018] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 684–699, 2018.

[Zhang *et al.*, 2021] Xuying Zhang, Xiaoshuai Sun, Yunpeng Luo, Jiayi Ji, Yiyi Zhou, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Rstnet: Captioning with adaptive attention on visual and non-visual words. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15465–15474, 2021.