EXPLORING DUAL STREAM GLOBAL INFORMATION FOR IMAGE CAPTIONING

Tiantao Xian[1], *Zhixin Li*[1,⋆], *Tianyu Chen*[1], *Huifang Ma*[2]

[1]Guangxi Key Lab of Multi-source Information Mining and Security
Guangxi Normal University, Guilin 541004, China
[2]College of Computer Science and Engineering,Northwest Normal University,Lanzhou 730070, China
⋆Corresponding Author. E-mail: lizx@gxnu.edu.cn

ABSTRACT

In recent years, image caption methods based on the encoder-decoder framework have made promising achievements, but most of them lack the exploitation of global information. In general, visual global information can provide more fine-grain details for recognizing small objects. On the other hand, the textual global information provides a coarse understanding of the visual scene. In this paper, we propose Dual Global Enhanced Transformer (DGET) to explicitly utilize both visual and textual global information. In encoding stages, we complement two visual features with different properties to obtain a global enhanced visual representation by a novel Global Enhanced Encoder (GEE). During decoding, we proposed Global Enhanced Decoder (GED) to utilize the textual global information explicitly. To validate our model, we conduct extensive experiments on the COCO image captioning dataset and achieve superior performance over many state-of-the-art methods.

Index Terms— Image Captioning, Transformer, Global Information, Attention mechanism

1. INTRODUCTION

Image captioning aims to describe the semantic content of an image in neural language, which has recently attracted extensive research attention. In the past few years, most image captioning model[1, 2, 3, 4, 5, 6] base on encoder-decoder framework. Attention mechanism is applied to prompt the decoder solely focus on the crucial region of image, and bring significant improvement on most evaluation metrics. Inspired by the developments of transformer architecture [7] in natural language processing, many recent works has studies its application in vision-language tasks. However, the current Transformer-base approaches [8, 9, 10, 11] still suffers from the following limitations: 1) In the encoding stage, the region features may fail to cover all objects in the image. It causes the problem of small object missing and incorrect relationship recognition between objects; 2) In the decoding stages, the prediction word at the current time step only depends on the previously generated word. In this case, the decoder is unable to capture sufficient semantic information, resulting in the generation of inaccurate words.

To solve the above problems, we propose Dual Global Enhanced Transformer (DGET), which consist of Global Enhanced Encoder (GEE) and Global Enhanced Decoder (GED). In the coding stage, GEE enables the complementary advantages of region-level and grid-level for a better visual representation. The relative geometry information of region feature was considered via Relation Enhanced Attention (REA). After that, the grid features are integrated into the region features by Visual Global Adaptively Attention (VGAA) module. The intuition is that the grid features contain more fine-grained information, while the region features contain object information. They can complement each other to achieve a better visual representation, which is essential to improve captioning performance. In GED, we first design a Context Encoder to encode the existing caption as context vector. Subsequently, we adaptively fuse the context vector into the decoder at each time step via Textual Global Adaptive Attention (TGAA). The intuition is that when humans understand a scene, they first find an approximate understanding based on experience, and then fine-tune it. On the other hand, the existing caption can serve as prior knowledge to guide the decoder to generate a more accurate caption. Hence, we regard the existing captions generated by the classical image captioning model as a rough understanding of the scene, and then incorporate it into decoder to mimic this process.

To sum up, our major contributions are summarized as follows: (1) We propose Global Enhanced Encoder to refine the representation of region features by leveraging grid features to provide more detailed information. At the same time, geometric information was combined to model complex spatial relationships of region features. (2) We design a Global

This work is supported by National Natural Science Foundation of China(Nos. 61966004, 61866004), Guangxi Natural Science Foundation(No. 2019GXNSFDA245018), Innovation Project of Guangxi Graduate Education, Guangxi "Bagui Scholar" Teams for Innovation and Research Project, Guangxi Talent Highland Project of Big Data Intelligence and Application, and Guangxi Collaborative Innovation Center of Multi-source Information Integration and Intelligent Processing.

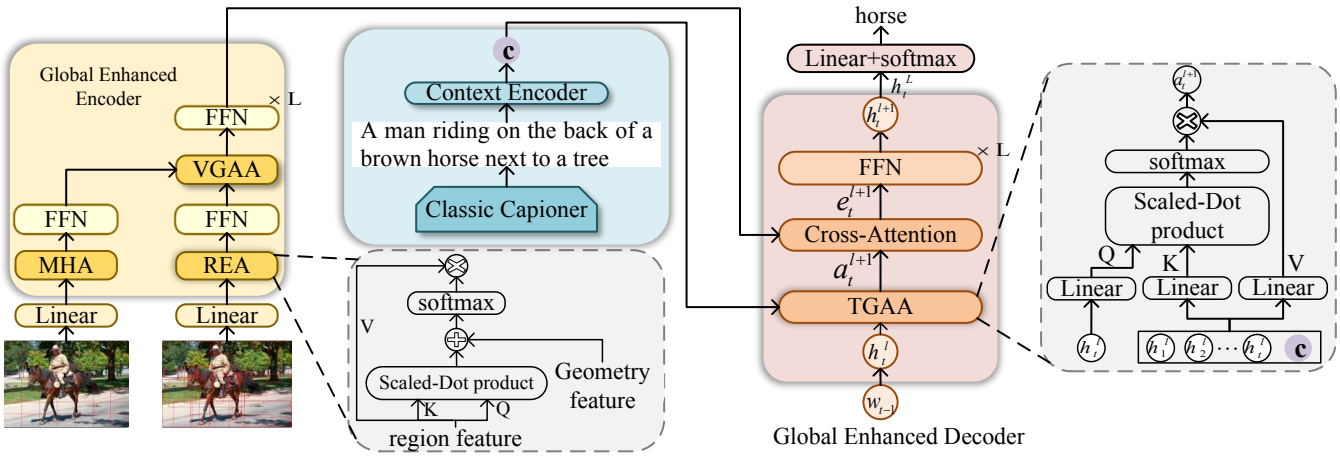

Fig. 1. The overall architecture of our DGET. Note that the residual connect and layer moralization are ignored for simplicity.

Enhanced Decoder, the decoder can adaptively attend the textual global information at each time step. In addition, we devise three different variants of Context Encoder to explore the impact of different representations of the existing caption on performance. (3) Extensive experiments on the benchmark COCO image captioning dataset are conducted to prove the usefulness of the proposed models. The experimental results demonstrate that our proposed method performs much better than other state-of-the-art methods.

2. PROPOSED APPROACH

2.1. Global Enhanced Encoder

Given an image, we first obtain the region features $V_r = \{r_1, r_2, \ldots, r_n\}$, where n is the number of region. Moreover, we consider the size and coordinates of the object's bounding box as geometric information, which represented as x, y, h, w (center coordinates, widths, and heights). We then obtain the grid features $V_g = \{g_1, g_2, \ldots, g_m\}$, where m denotes the number of the grid. To adapt the feature dimensionality to the encoder, the visual features V_r and V_g is first fed into fully-connected linear layer respectively, then we get projected features ${V_r}^0 \in \mathbb{R}^{n\times d}$ and ${V_g}^0 \in \mathbb{R}^{m\times d}$.

We take the calculation process of $(l+1)$-th $(0 \le l \le L)$ layer of GEE as an example. we first model intra-level relationships of two kinds of features by Multi-Head Attention (MHA) module and Relation Enhanced Attention (REA), respectively:

$$O_g^{l+1} = \text{Add\&Norm}(\text{MHA}(v_g^l, v_g^l, v_g^l)) \quad (1)$$

$$O_r^{l+1} = \text{Add\&Norm}(\text{REA}(v_r^l, v_r^l, v_r^l, G^l)) \quad (2)$$

Then, we employ two independent FFN modules after the MHA module and REA module:

$$V_g^{l+1} = \text{Add\&Norm}(\text{FFN}(O_g^{l+1})) \quad (3)$$

$$M_r^{l+1} = \text{Add\&Norm}(\text{FFN}(O_r^{l+1})) \quad (4)$$

The region-level futures may not cover the entire image, which inevitably missing small objects and scenes details. However, these shortcomings are the advantage of grid-level features, which in contrast, cover all the content of a given image. It is natural that integrate the grid features into the region features to obtain better visual representation. We complete this process via our proposed Visual Global Adaptively Attention (VGAA) module.

$$\overline{V}_r^{l+1} = \text{Add\&Norm}(\text{VGAA}(M_r^{l+1}, V_g^{l+1}, V_g^{l+1})) \quad (5)$$

$$V_r^{l+1} = \text{Add\&Norm}(\text{FFN}(\overline{V}_r^{l+1})) \quad (6)$$

The detail of VGAA module is described in the next section. After multi-layer encoding, the region features are fed into decoder layers.

2.1.1. Relation Enhanced Attention

The permutation invariance of the Transformer structure is suitable for using the region features. It can model the positional relationship without being affected by order of input. We calculate the relative position and size of the bounding boxes between two objects i and j, which denote as $\omega(i, j)$:

$$\left(\log(\frac{|x_i - x_j|}{w_i}), \log(\frac{|y_i - y_j|}{h_i}), \log(\frac{w_i}{w_j}), \log(\frac{h_i}{h_j})\right)^{\mathrm{T}} \quad (7)$$

where x_i, y_i, w_i, h_i are the center coordinate, width, and height of box i, respectively. The geometric attention weights are then calculated as $G_{ij} = \text{ReLU}(\text{FC}(\text{Emb}(\omega)))$.

We take the effect of G into the calculation of self-attention as:

$$\text{REA}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}} + \log(G))V \quad (8)$$

Note that the Emb method is following Vaswani et al.[7], which embed ω to a high-dimensional representation.

2.1.2. Visual Global Adaptively Attention

We take the grid-level representation vector as key and value and regard the region-level representation vector as the query in the multi-head attention module. In this way, the region attend to each fragment of grid to realize the complementary of visual information from two different sources feature:

$$\bar{V}_r^{l+1} = \mathrm{MHA}(M_r^{l+1}, V_g^{l+1}, V_g^{l+1}) \tag{9}$$

2.2. Global Enhanced Decoder

In the decoding phase, we denote $w_t \in \mathbb{R}^d$ as the vector representation of the t-th word, which is the sum of word embedding and positional encoding. Therefore, the input matrix for time step t is $W_{t-1} = \{w_0, w_1, \ldots, w_{t-1}\} \in \mathbb{R}^{d\times t}$. We denote the embedded existing caption generated by the classic captioning model as $P = \{p_1, p_2, \ldots, p_k\} \in \mathbb{R}^{d\times t}$, where k is the length of existing caption. Before decoding, we encode the existing caption via the proposed Context Encoder to obtain a context vector:

$$c = \mathrm{ContextEncoder}(P) \tag{10}$$

where $c \in \mathbb{R}^d$ is the context vector which can be view as sentence embedding of existing caption. The detail of Context Encoder is described in the next subsection.

Suppose that the decoder is in the generation process of the t time step, the input of $(l+1)$-th layer is $H_t^l = \{h_1^l, h_2^l, \ldots, h_t^l\} \in \mathbb{R}^{d\times t}$ and context vector c. They are fed into our proposed Textual Global Adaptive Attention (TGAA) module to model the relationship for the left-hand sequences and adaptively capture global information simultaneously. Subsequently, we denote the output of TGAA as a_t^{l+1}, and passed it into the multi-head cross-attention to incorporate with visual feature. Finally, we fed the output of multi-head cross-attention (denote as e_t^{l+1}) into a feed-forward neural network (FFN).

$$a_t^{l+1} = \mathrm{Add\&Norm}(\mathrm{TGAA}(h_t^l, H_t^l, H_t^l, c)) \tag{11}$$

$$e_t^{l+1} = \mathrm{Add\&Norm}(\mathrm{MHA}(a_t^{l+1}, V_r^L, V_r^L)) \tag{12}$$

$$h_t^{l+1} = \mathrm{Add\&Norm}(\mathrm{FFN}(e_t^{l+1})) \tag{13}$$

The detail of TGAA is described in the following subsection.

2.2.1. Context Encoder

LSTM-LAST. LSTM is widely used in sequence data encoding, which incorporates suitable information from the current time step via the input gate and forgets useless information from the previous step via the forget gate. We take the last hidden state of LSTM output as a context vector to represent the sentence:

$$c = h_k = \mathrm{LSTM}(p_1, \ldots, p_k) \tag{14}$$

BiLSTM-MAX. A more sophisticated method is to use the Bidirectional LSTM, in which the sentences are encoded forward and backward with LSTM, respectively. We select the maximum value over each dimension of the set of hidden units as context vector:

$$\begin{aligned} h &= \mathrm{BiLSTM}(p_1, \ldots p_k) \\ c &= \mathrm{MaxPooling}(h) \end{aligned} \tag{15}$$

Transformer-Encoder. We also attempt to embed the sentences through the self-attention module. The encoder contains six stacked identical layers. We add special token CLS to each layer and regard the CLS of the last encode layer as the context vector. We first calculating the $\mathrm{CLS} = (\sum p_i)/k$. Then, we set $P_g = (P; \mathrm{CLS})$, and pass it into the Transformer Encoder structure:

$$\begin{aligned} \bar{P}_g^{l+1} &= \mathrm{Add\&Norm}(\mathrm{MHA}(P_g^l)) \\ P_g^{l+1} &= \mathrm{Add\&Norm}(\mathrm{FFN}(\bar{P}_g^{l+1})) \end{aligned} \tag{16}$$

Finally, we take the token CLS on the output of the last encoder layer as context vector c.

2.2.2. Textual Global Adaptive Attention

The efficient method uses the multi-head attention for fusion, which naturally fuses the local representation and the global representation by taking a weighted sum of hidden state H_t^l and context vector c in the t-th time step. We set $H_{t,g}^l = (H_t^l; c)$. The calculation of TGAA as below:

$$a_t^{l+1} = \mathrm{MHA}(h_t^l, H_{t,g}^l, H_{t,g}^l) \tag{17}$$

3. EXPERIMENTS

All the experiments are conducted on the most popular benchmark dataset of image captioning, i.e., COCO. In offline evaluation, we adopt the Karpathy splits [12]. This split contains 123,287 images with available annotation, including 113,287 images for training, and 5000 images for validation and 5,000 for testing. The online evaluation is done on the COCO test server, for which ground truth annotations are not publicly available.

For images features, we utilize the bottom-up features [3] as region features. Each region is represented as a 2048-dimensional vector. We use employ pre-trained ResNet-152 to extract grid features with the size of $7 \times 7 \times 2048$. For existing caption, we use our reproduced Up-Down captioner [3]. we set d to 512 and the number of head to 8. The number of layers for encoder and decoder is set to 6, k is set 5 and batch size to 50. In the cross-entropy(XE) pre-training stage, we following the learning rate scheduling strategy with a warmup equal to 20, 000 iterations. After the 20-epoch XE pre-training stage, we start to optimize our model with CIDEr reward following [13] with 5×10^{-6} learning rate.

Table 1. Performance comparisons with other advanced single models on the official COCO test server.

Model	B-4		M		R		C	
	c5	c40	c5	c40	c5	c40	c5	c40
Up-Down[3]	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
GCN-LSTM[19]	38.7	69.7	28.5	37.6	58.5	73.4	125.3	126.5
SGAE[20]	37.8	68.7	28.1	37.0	58.2	73.1	122.7	125.5
ETA[10]	38.9	70.2	28.6	38.0	58.6	73.9	122.1	124.4
NG-SAN[21]	38.8	70.2	**29.0**	**38.4**	58.7	74.0	126.3	128.6
Ours (w/LSTM-L)	**39.2**	70.4	28.8	38.1	58.8	74.1	**126.6**	128.5
Ours (w/BiLSTM-M)	**39.2**	70.4	28.9	38.2	58.8	74.0	**126.6**	128.2
Ours (w/Trans-Enc)	39.1	**71.2**	28.9	**38.4**	**58.9**	**74.4**	126.3	**129.2**

Table 2. Performance comparisons on Karpathy test split. † means results from our reproductions.

Model	B-1	B-4	M	R	C	S
Up-Down[3]	79.8	36.3	27.7	56.9	120.1	21.4
Up-Down†[3]	79.4	36.7	27.9	57.6	122.7	21.4
GCN-LSTM[19]	80.5	38.2	28.5	58.3	127.6	22.0
SGAE[20]	80.8	38.4	28.4	58.6	127.8	22.1
ORT[22]	80.5	38.6	28.7	58.4	128.3	22.6
ETA[10]	**81.5**	39.3	28.8	58.9	126.6	22.7
M2[8]	80.8	39.1	29.2	58.6	131.2	22.6
NG-SAN[21]	-	39.9	**29.3**	59.2	132.1	**23.3**
GET[9]	**81.5**	39.5	**29.3**	58.9	131.6	22.8
Ours (w/LSTM-L)	81.2	39.9	29.2	59.2	132.0	23.1
Ours (w/BiLSTM-M)	81.3	40.1	29.2	59.2	132.1	23.1
Ours (w/Trans-Enc)	81.3	**40.3**	29.2	**59.4**	**132.4**	**23.3**

Five evaluation metrics, i.e., BLEU [14], METEOR[15], ROUGE-L [16], CIDEr [17], and SPICE [18], are simultaneously utilized to evaluate our model.

3.1. Comparison With State-Of-The-Art Methods

Offline Evaluation. Table 2 shows the performance of the sota models and our approach on the offline COCO Karpathy [12] test split. The compared models include: Up-Down [3], GCN-LSTM [19], SGAE [20], ORT [22], ETA[10], M2 [8], NG-SAN [21] and GET [9]. The result indicates that our method surpasses all other approaches in terms of BLEU-4, ROUGE-L, CIDEr and SPICE, and achieves competitive performance on BLEU-1 and METEOR compared to the SOTA approach. In particular, it advances the current state-of-the-art on CIDEr and BLEU-4 by 0.3 and 0.4 points, respectively.
Online Evaluation. Table 1 reports the performance of our proposed model and other top-ranking published models on the COCO test server. Note that the ensemble model always has better performance. For fair comparison, we use the single model to compare with the published state-of-the-art single models[3, 19, 10, 20, 21]. As can be seen, compared with the published methods, our single model significantly outperforms all the other methods in terms of BLEU-4(c5, c40), ROUGE(c5, c40) and CIDEr(c5, c40), and achieves competitive performance on METEOR(c5, c40).

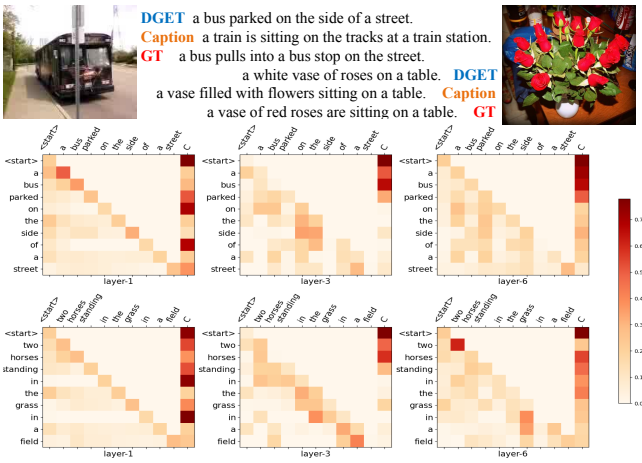

Fig. 2. Visualizations of the 1st, 3rd and 6th attention maps of TGAA module. c is the context vector.

3.2. Attention Analysis of the Decoder

As shown in Figure 2, we set the existing caption of low quality (i.e., the first example) and high quality (i.e., the second example). We find that the largest attention value in Layer-1 almost appears on the context vector, which indicates that our GED tends to capture the semantic content of the context vector at a shallow level. In contrast, the attention weights in layer-3 and layer-6 do not focus heavily on the context vector, indicating that our GED captures other semantic information at a deeper level. Besides, the attention weight to the context vector decreases significantly with the increase of time step, which indicates that the model gradually does not need the guidance of the textual global information when the semantic content of the generated sentences tends to be complete. By comparing Example 1 with Examples 2, we can observe that the attention values for context vectors encoded by low-quality are significantly lower than for high-quality ones, which shows that our model does not blindly follow the Context vector. These further demonstrate the advantage of our model.

4. CONCLUSIONS

In this paper, we propose Dual Global Enhanced Transformer (DGET) for image captioning, which consist of Global Enhanced Encoder (GEE) and Global Enhanced Decoder (GED). GEE use grid feature to provide visual global information for region feature to generate more comprehensive visual representation. GED adaptively fuses textual global information into the decoder to obtain more semantic representation at each time step. We also proposed three different Context Encoders to explore the existing caption better. Extensive results demonstrate the superiority of our approach that achieves a new sota on both offline and online evaluation.

5. REFERENCES

[1] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.

[2] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.

[3] 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*, 2018, pp. 6077–6086.

[4] Feicheng Huang, Zhixin Li, Haiyang Wei, Canlong Zhang, and Huifang Ma, "Boost image captioning with knowledge reasoning," *Machine Learning*, vol. 109, no. 12, pp. 2313–2332, 2020.

[5] Haiyang Wei, Zhixin Li, Canlong Zhang, and Huifang Ma, "The synergy of double attention: Combine sentence-level and word-level attention for image captioning," *Computer Vision and Image Understanding*, vol. 201, pp. 103068, 2020.

[6] Haiyang Wei, Zhixin Li, Feicheng Huang, Canlong Zhang, Huifang Ma, and Zhongzhi Shi, "Integrating scene semantic knowledge into image captioning," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 17, no. 2, pp. 1–22, 2021.

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

[8] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara, "Meshed-memory transformer for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10578–10587.

[9] Jiayi Ji, Yunpeng Luo, Xiaoshuai Sun, Fuhai Chen, Gen Luo, Yongjian Wu, Yue Gao, and Rongrong Ji, "Improving image captioning by leveraging intra-and inter-layer global representation in transformer network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 1655–1663.

[10] Guang Li, Linchao Zhu, Ping Liu, and Yi Yang, "Entangled transformer for image captioning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8928–8937.

[11] Tianyu Chen, Zhixin Li, Tiantao Xian, Canlong Zhang, and Huifang Ma, "Relation also need attention: Integrating relation information into image captioning," in *Asian Conference on Machine Learning*. PMLR, 2021, pp. 1537–1552.

[12] Andrej Karpathy and Li Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.

[13] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel, "Self-critical sequence training for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7008–7024.

[14] 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*, 2002, pp. 311–318.

[15] 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*, 2005, pp. 65–72.

[16] Chin-Yew Lin, "Rouge: A package for automatic evaluation of summaries," in *Proceedings of the ACL Workshop on Text Summarization Branches Out*, 2004, pp. 74–81.

[17] 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*, 2015, pp. 4566–4575.

[18] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould, "Spice: Semantic propositional image caption evaluation," in *European conference on computer vision*. Springer, 2016, pp. 382–398.

[19] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei, "Exploring visual relationship for image captioning," in *Proceedings of the European conference on computer vision*, 2018, pp. 684–699.

[20] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai, "Auto-encoding scene graphs for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10685–10694.

[21] 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 Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10327–10336.

[22] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares, "Image captioning: transforming objects into words," in *Proceedings of the International Conference on Neural Information Processing Systems*, 2019, pp. 11137–11147.