MIXED KNOWLEDGE RELATION TRANSFORMER FOR IMAGE CAPTIONING

*Tianyu Chen, Zhixin Li**, *Jiahui Wei, Tiantao Xian*

Guangxi Key Lab of Multi-source Information Mining and Security,
Guangxi Normal University, Guilin 541004, China
*Corresponding Author. E-mai:lizx@gxnu.edu.cn

ABSTRACT

Internal relationship of image objects has contributed significantly to the development of image captioning, especially when combined with Transformer architecture. Most of these methods only calculate the relationship between entities and ignore the information between entities and background. Besides, the way of exploring the relational information inside the image can also be extended. In this paper, we continually explore the relationship between objects from both internal and external perspectives, and embed the vital image global information into the internal relationship module. To validate the effectiveness of our model, we conduct extensive experiments on the most popular MSCOCO dataset, and achieve state-of-the-art performance on both online and offline test sets.

Index Terms— image captioning, external knowledge, object relation

1. INTRODUCTION

Accurately capturing the semantic information of the main objects in the image and describing the relationship between them is a critical task in image captioning. Recently the Visual Transformer [1, 2, 3, 4] models have been a great success because of the excellent modeling and exploitation of the internal image relations using self-attention [5]. On this basis of region-level features [6], many advanced models [1] [2] have explored the correlation between objects and calculated the attention weights in the way of visual and geometric relationships.

Despite the great success in the previous exploration of object internal connections, the perspective of modeling relations is still limited to the interior of the image. Only considering internal relations will bring some visual interference. As illustrated in Fig. 1 (a) and (b), the green regions are some top-5 attended objects by the red regions(from deep to shadow), in Transformer(left), some green regions are calculated to have high similarity to the red region because of the similar color, shape, and close relative position, but are not top semantically related. For example, the *white cat* with the *white pant* in (a), the *man* on the far right with the *American flag* in (b). However, humans are able to discard this visual interference because of the ability to convert visual information into semantic information and consider the global semantic coherency based on their empirical knowledge. This inspires us to explore how to incorporate this kind of knowledge and global perspective into image captioning model to imitate human reasoning procedure.

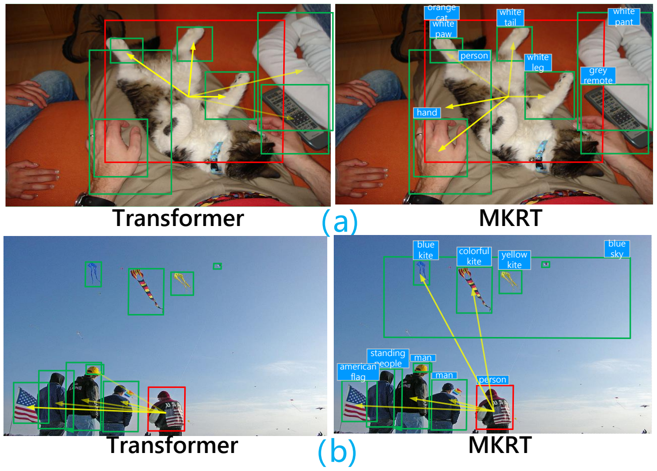


Fig. 1. Examples of visual interference problem and our MKRT improvements.

According to human commonsense, if two objects are highly semantically related, the probability of them appearing in the same image is also high. But this semantic guidance cannot be obtained from inside the image. In this paper, we utilize external knowledge obtained from the statistical method and propose a Mixed Knowledge Relation Transformer(MKRT). What's more, the detected objects usually contain some important background information, such as the *sky* in Fig. 1 (b). In order to compensate the global and background information, we propose a Global Enhanced Module(GEM) to mine the objects-global association.

*This work is supported by National Natural Science Foundation of China(Nos. 61966004, 61866004), Guangxi Natural Science Foundation(No. 2019GXNSFDA245018), 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.

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

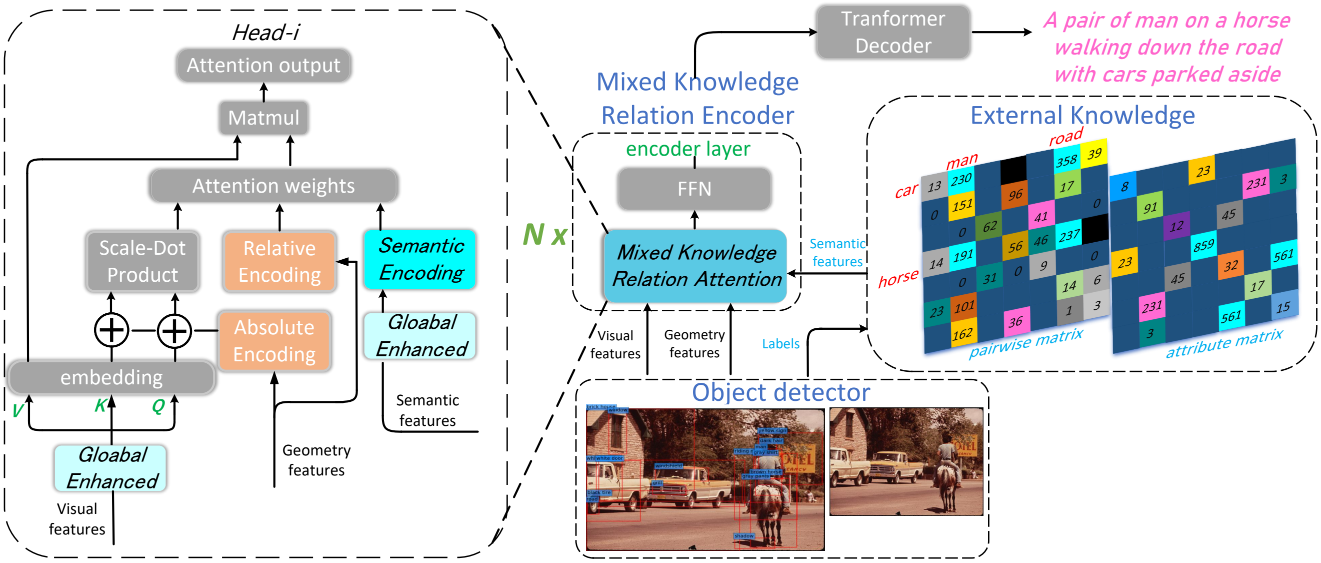


Fig. 2. Overview of the proposed Mixed Knowledge Relation Transformer.

To sum up, our major contributions are itemized below:

- To alleviate visual interference in the traditional relation modules, we use external knowledge to incorporate semantic information between objects into relation modeling and propose the Mixed Knowledge Relation Transformer
- We extract a comprehensive global representation of the image in GEM, which can further mine the object-global relation, and successfully embed it into the self-attention paradigm.
- Extensive experiments demonstrate that our MKRT can significantly improve performance. When GRM is incorporated, the improvement is even more pronounced and achieves new state-of-the-art performance.

2. MIXED KNOWLEDGE RELATION TRANSFORMER

Our MKRT model consists of two main parts: the Mixed Knowledge Relation Attention(MKRA) and the Global Enhanced Module(GEM). As shown in Fig. 2, the overall architecture follows the conventional visual Transformer paradigm. We will introduce the key components of our MKRT in the following.

2.1. The External Knowledge

The external knowledge extracted from Visual Genome(VG) dataset [7] includes the pairwise matrix, which presents the category co-occurrence probability, and the attribute matrix used to present the attribute similarity between objects. The synsets is set as matrix index. The value $\boldsymbol{M}^{P}_{i,j}$ of the pairwise matrix $\boldsymbol{M}^{P}_{C,C}$ represents the frequency when category-i appears in an image, category-j also appears, but this value is not equal to the $\boldsymbol{M}^{P}_{j,i}$.

For the attribute knowledge, we count a C×K frequency distribution for each category-attribute pair, then the Jensen-Shannon(JS) divergence [8] between the frequency distribution F_i of class-i and F_j of class-j is measured as the value of the attribute matrix $\boldsymbol{M}^{A}_{i,j}$, $\boldsymbol{M}^{A}_{i,j}$=$JS(F_i \| F_j)$. The $\boldsymbol{M}^{A}_{C,C}$ is symmetric. The higher the value in both matrices, the higher the semantic similarity between categories.

2.2. Mixed Knowledge Relation Attention

Besides the visual relation Ω_A calculated by the conventional self-attention, the relative geometry relation is also considered. The relative geometric relation is calculated by:

$$\omega_G^{mn} = ReLU(Emb(\lambda)W_G \tag{1}$$

where Emb(•) conducts a high-dimensional embedding and λ is the relative position vector.

Next, the absolute geometric relation is integrated into the visual relation Ω_A and then embedded together into $\hat{\omega}_A^{mn}$, which has the same dimension with ω_G^{mn}:

$$\hat{\omega}_A^{mn} = \frac{(Q + pos_q)(K + pos_k)^T}{\sqrt{d_k}} \tag{2}$$

In our MKRA, the important semantic information are considered in the attention weights calculating, as illustrated in Fig. 2, the category and attribute labels derived from the object detector are used as the index of the external knowledge matrices. A hyper parameter α is set to measure the reasonable proportion of two matrix values. The semantic relation between *m* and *n* is expressed as follows:

$$\Omega^{S}_{mn} = \alpha M^{P}_{m,n} + (1-\alpha) M^{A}_{m,n} \tag{3}$$

We embed the 1-d vector Ω^{S}_{mn} into the same dimension with ω^{mn}_{G}:

$$\omega^{mn}_{S} = ReLU(Emb(\Omega^{S}_{mn})W_S) \tag{4}$$

$\boldsymbol{W_S}$ is a learned parameter matrix. We fuse the three kinds of relation weights together and the combined attention weights $\hat{\omega}^{mn}$ is:

$$\hat{\omega}^{mn} = \frac{log(\omega^{mn}_{G}) + \hat{\omega}^{mn}_{A} + log(\omega^{mn}_{S})}{\sum_{l=1}^{N} log(\omega^{ml}_{G}) + (\hat{\omega}^{mn}_{A}) + log(\omega^{ml}_{S})} \tag{5}$$

Finally, the softmax is utilized to normalize weights and calculate the outputs of MKRA. The Multi-Head mechanism is also combined with our MKRA the same as Trasnformer. The Head-i MKRA can be formalized as:

$$H_i = MKRA(Q, K, V, pos_q, pos_k, \Omega^{S}, \Omega_A) \tag{6}$$

$$= softmax(\hat{\omega}^{mn}_{A} + log(\Omega_A) + log(\Omega_S))V \tag{7}$$

2.3. Global Enhanced Module

Besides the intra-objects relation, the objects-global is also valued. The detected object appearance features are present as $[V_1,...,V_N]$, we average them into a global feature V_g

$$V_g = \frac{1}{N}\sum_{i=1}^{N} v_i \tag{8}$$

The corresponding semantic relation corresponding to V_g is calculated through the pairwise and attribute matrices. We average the values of the corresponding rows or columns in two matrices. Suppose the coordinate of the global vector is *g* and an object's synset is *m* in VG dataset, the semantic relation between this object to global is:

$$M_{m,g} = \frac{1}{N}\sum_{i=1}^{N} M_{m,i}, \quad M_{g,m} = \frac{1}{N}\sum_{i=1}^{N} M_{i,m} \tag{9}$$

Since the global vector does not have a clear location, we set its absolute position to 0 and its relative position to 1. Now we get all values of the global vector in Equation 5, it can be computed just like any other region-level vector in MKRA.

2.4. Objectives

Given the ground truth sequence $y^{*}_{1:T}$, the parameter of model is denoted as θ, the conventional cross-entropy(XE) loss is:

$$L_{XE} = -\sum_{t=1}^{T} log(p_\theta(y^{*}_{t}|y^{*}_{1:t-1})) \tag{10}$$

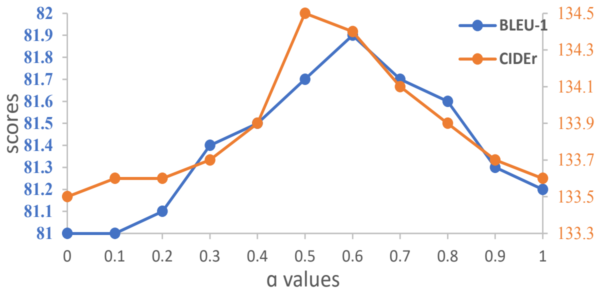


Fig. 3. Experiments on different α setting and the corresponding performance.

Besides the *XE* loss, we also optimize the non-differentiable CIDEr-D score by Self-critical Sequence Training [9] following [3]:

$$\nabla_\theta L_{RL}(\theta) = -\frac{1}{k}\sum_{i=1}^{k}(r(y^{i}_{1:T}) - b)\nabla_\theta log p_\theta(y^{i}_{1:T}) \tag{11}$$

where *k* is the beam size, *r* is the CIDEr-D score reward, and b=$(\sum_i r(y^{i}_{1:T}))/k$ is the reward baseline to reduce the high-variance caused by sampling.

3. EXPERIMENTS

3.1. Result and Analysis

We conduct our experiments with standard evaluation metrics like BLEU [10] and so on. The *c* is 3000 and *K* is 200 in exernal knowledge. The best setting for our hyper parameter is discussed in Fig. 3, refering to the BLEU-1 and CIDEr [11] scores, we set our α to 0.6 for balancing the category co-occurrence probability and attribute similarity. First of all to better verify the effectiveness of our MKRA module, we choose ORT in [1] and DLCT [2] as our baseline. For fair comparison, we train our model for 18-epoch for DLCT baseline and 30-epoch for ORT baseline in the XE training phase with ResNet-101 [20] as our backbone, the results are shown in Table 1. As we can see, our MKRA module without GEM can effectively improve the performance of ORT and DLCT. At the same time, our model achieve the new state-of-the-art performance in CIDEr Optimizationn stage compared with other advanced models. We further ensemble our MKRA and GEM modules in MKRT and submit the generated captions into the online test server, as can be seen in Table 2, our MKRT is very competitive and able to achieve comparable results to other models in most of metrics.

3.2. Ablation Study

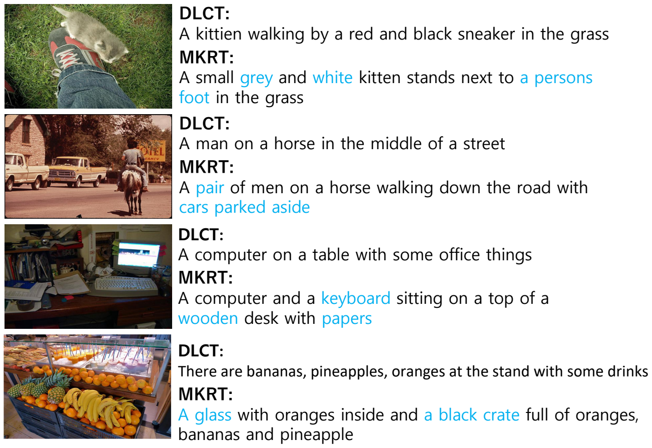
Several ablative studies are conduct to quantify the contribution of each design. The performance of each component and their combination is shown in Table. 3. From the table, we can see that the MKRA module has a more obvious improve-

Table 1. Performance comparisons on offline COCO Karpathy test split with other advanced models, where B-N, M, R, C and S are short for BLEU-N, METEOR, ROUGE-L, CIDEr and SPICE scores.

	Cross Entropy						CIDEr Optimization					
Model	B-1	B-4	M	R	C	S	B-1	B-4	M	R	C	S
DAIC [12]	73.7	34.2	26.4	54.8	106.2	-	77.6	35.4	26.7	56.5	116.8	-
HEK [13]	-	-	-	-	-	-	79.3	37.3	27.3	57.4	121.2	-
VASS [14]	76.9	36.5	27.9	56.6	114.0	20.8	80.5	38.9	28.3	58.8	126.7	21.7
AoANet [15]	77.4	37.2	28.4	57.5	119.8	21.3	80.2	38.9	29.2	58.8	129.8	22.4
X-Transformer [16]	77.3	37.0	28.7	**57.5**	**120.0**	21.8	80.9	39.7	29.5	59.1	132.8	23.4
M2 Transformer [3]	-	-	-	-	-	-	80.8	39.1	29.2	58.6	131.2	22.6
GET [17]	-	-	-	-	-	-	81.5	38.5	29.3	58.9	131.6	22.8
ORT [1]	75.1	34.0	27.7	55.7	115.0	20.9	80.5	38.6	28.7	58.4	127.8	22.1
Our(ORT+MKRA)	**79.3**	**37.6**	**28.9**	57.3	119.2	**21.9**	81.3	39.1	29.4	59.2	131.3	22.6
DLCT [2]	-	-	-	-	-	-	81.4	39.8	29.5	59.1	133.8	23.0
Our(DLCT+MKRA)	-	-	-	-	-	-	**81.9**	**40.1**	**29.9**	**59.9**	**134.5**	**23.5**

Table 2. Leaderboard of various methods on the online COCO test server.

	B-1		B-2		B-3		B-4		M		R		C	
Model	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
SGAE [18]	81.0	95.3	65.6	89.5	50.7	80.4	38.5	69.7	28.2	37.2	58.6	73.6	123.8	126.5
AoA [15]	81.0	95.0	65.8	89.6	51.4	81.3	39.4	71.2	29.1	38.5	58.9	74.5	126.9	129.6
HIP [19]	81.6	95.9	66.2	90.4	51.5	81.6	39.3	71.0	28.8	38.1	59.0	74.1	127.9	130.2
M2 [3]	81.6	96.0	66.4	90.8	51.8	82.7	39.7	72.8	29.4	39.0	59.2	74.8	129.3	132.1
X-Transformer [16]	81.3	95.4	66.3	90.0	51.9	81.7	39.9	71.8	29.5	39.0	59.3	74.9	129.3	131.4
DLCT [2]	82.0	96.2	66.9	91.0	52.3	83.0	40.2	73.2	**29.5**	**39.1**	**59.4**	**74.8**	131.0	133.4
Our(MKRT)	**82.1**	**96.4**	**67.2**	**91.3**	**52.6**	**83.2**	**40.5**	**73.5**	29.4	39.0	59.1	74.6	**131.2**	**133.5**



DLCT:
A kittien walking by a red and black sneaker in the grass
MKRT:
A small grey and white kitten stands next to a persons foot in the grass
DLCT:
A man on a horse in the middle of a street
MKRT:
A pair of men on a horse walking down the road with cars parked aside
DLCT:
A computer on a table with some office things
MKRT:
A computer and a keyboard sitting on a top of a wooden desk with papers
DLCT:
There are bananas, pineapples, oranges at the stand with some drinks
MKRT:
A glass with oranges inside and a black crate full of oranges, bananas and pineapple

Fig. 4. The captions generated by our model and the baseline

ment than GEM for baseline model, since the semantic relation plays a major role. The improvement is more obvious when the two modules are combined together in MKRT.

3.3. Qualitative Analysis

The effect of semantic guidance is intuitively shown in Fig. 4. As we can seen many word pairs with high semantic relation are accurately captured, which indicates that our model can understand the image content at a deeper level and generate more anthropomorphic descriptions.

Table 3. Settings and results of ablation studies.

Models	B-1	B-4	M	R	C
base(DLCT)	81.4	39.8	29.5	59.1	133.8
base+MKRA	81.9	40.1	29.9	59.9	134.5
base+GEM	81.6	39.8	29.7	59.5	134.1
base+MKRA+GEM	**82.1**	**40.2**	**31.1**	**60.2**	**134.8**

4. CONCLUSION

Extensive experiments have demonstrated the effectiveness of semantic coherency and object-global relation. Better combination with image grid features are considerd in the future. We represent the semantic information between targets by statistically extracted external knowledge and embed it into the previous relation module. Meanwhile, we propose a method to represent global visual and semantic information to further mine the connection between targets and the global. Both designs are experimentally shown to give a boost to the baseline models.

5. REFERENCES

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, 2017.

[8] ML Menéndez, JA Pardo, L Pardo, and MC Pardo, "The jensen-shannon divergence," *Journal of the Franklin Institute*, vol. 334, no. 2, pp. 307–318, 1997.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei, "Attention on attention for image captioning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 4634–4643.

[16] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei, "X-linear attention networks for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10971–10980.

[17] 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.

[18] 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.

[19] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei, "Hierarchy parsing for image captioning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 2621–2629.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.