International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

VTKG: A Vision Transformer Model with Integration of Knowledge Graph for Enhanced Image Captioning

Ms. Yugandhara A. Thakare[1], Dr. K. H. Walse[2], Dr. Mohammad Atique[3]
[1]P.G. Department of Computer Science & Engineering, Sant Gadge Baba Amravati University, Amravati, India
yugathakare@gmail.com
[2]Sant Bhagwan Baba Kala Mahavidyalaya, Sindkhed Raja, India
kwalse1234@gmail.com
[3]P.G. Department of Computer Science and Engineering, Sant Gadge Baba Amravati University, Amravati, India

Abstract—The Transformer model has exhibited impressive results in machine translation tasks. In this research, we utilize the Transformer model to improve the performance of image captioning. In this paper, we tackle the image captioning task from a novel sequence-to-sequence perspective and present VTKG, a VisionTransformer model with integrated Knowledge Graph, a comprehensive Transformer network that substitutes the CNN in the encoder section with a convolution-free Transformer encoder. Subsequently, to enhance the generation of meaningful captions and address the issue of mispredictions, we introduce a novel approach to integrate common-sense knowledge extracted from a knowledge graph. This has significantly improved the overall adaptability of our captioning model. Through the amalgamation of the previously mentioned strategies, we attain exceptional performance on multiple established evaluation metrics, outperforming existing benchmarks. Experimental results demonstrate a 1.32%, 1.7%, 1.25%, 1.14%, 2.8% and 2.5% improvement in Blue-1, Bluu-2, Blue-4, Metor, Rough-L and CIDEr score respectively when compared to state-of-the-art methods.

Keywords-Knowledge graph, Transformer model, Vision Transformer, image captioning.

I. INTRODUCTION

Image caption generation has witnessed significant advancements by integrating commonsense knowledge and reasoning abilities into the process. This approach aims to enhance the quality and coherence of generated captions by leveraging our innate understanding of the world and incorporating logical inference. The combination of visual perception and commonsense reasoning allows AI systems to produce captions that are not only visually grounded but also contextually and semantically meaningful.

To incorporate commonsense knowledge, researchers have relied on resources such as ConceptNet, WordNet, and Cyc, which provide models with a rich background of knowledge about objects, actions, relationships, and events. This knowledge enriches the understanding of visual scenes and aids in generating captions that demonstrate a deeper understanding of the depicted content [1].

The integration of reasoning mechanisms into image captioning has been explored to go beyond surface-level descriptions. Techniques such as logical inference, spatial reasoning, temporal reasoning, and causal reasoning have been employed to generate captions that exhibit logical coherence and capture causal relationships. These mechanisms enable models to generate more accurate and human-like captions ([2], [3]).

Recent works have demonstrated the effectiveness of incorporating commonsense and reasoning abilities into image captioning. [1] Proposed a method that incorporates commonsense cues into the attention mechanism of the model, resulting in more coherent and contextually appropriate captions. [2] Developed a reasoning-enhanced image captioning model that leverages explicit logical inference to generate captions reflecting a deeper understanding of the depicted scene. [3] Explored the integration of spatial and temporal reasoning, resulting in captions that capture the spatial layout of objects and the temporal dynamics of events in the image.

In conclusion, the integration of commonsense knowledge and reasoning abilities holds great promise for advancing image caption generation. By leveraging background knowledge and incorporating reasoning mechanisms, AI systems can generate captions that not only accurately describe visual content but also demonstrate a deeper understanding of the scene and the relationships between objects and events.

II. RELATED WORK

A. Commonsense knowledge

Image caption generation has seen significant advancements in recent years, with a growing focus on integrating

commonsense knowledge and reasoning abilities into the process.

Hu et al. (2019) proposed a language-conditioned graph network for relational reasoning in image captioning. The model utilized commonsense cues encoded in ConceptNet to guide the attention mechanism, resulting in more coherent and contextually appropriate captions [1].

Zhang et al. (2021) developed a reasoning-enhanced approach for image captioning. The model integrated explicit logical inference into the caption generation process, enabling the generation of captions that reflected a deeper understanding of the depicted scene and captured causal relationships between objects and events [2].

Li et al. (2020) explored the integration of spatial and temporal reasoning into image captioning. The model employed spatial attention mechanisms and temporal reasoning modules to generate captions that captured the spatial layout of objects and the temporal dynamics of events in the image [3].

Zhang et al. (2019) introduced a semantic compositional network for image captioning. The model incorporated commonsense knowledge from ConceptNet and utilized compositional reasoning to generate captions that captured the semantics and relationships between objects in the image[4].

Yang et al. (2020) proposed a visual commonsense reasoning framework for image captioning. The model leveraged pretrained language models and commonsense knowledge bases to enable reasoning over visual scenes, resulting in captions that demonstrated a deeper understanding of the depicted content [5].

Gan et al. (2017) introduced a novel approach called "Semantic Compositional Networks" for image captioning. This approach leveraged commonsense knowledge and employed compositional reasoning to generate captions that captured the semantics and relationships between objects in the image [6].

Hwang et al. (2018) proposed a method called "Improving Image Captioning by Leveraging Knowledge Graphs." The model utilized ConceptNet, a large-scale commonsense knowledge graph, to enhance the caption generation process. By integrating the knowledge graph, the model generated captions that demonstrated a deeper understanding of the visual content [7].

Chen et al. (2019) introduced "Image Captioning with Compositional Hierarchical Tree Structures." The model employed a hierarchical tree structure to capture compositional reasoning and generate captions. By incorporating commonsense knowledge, the model produced captions that exhibited logical coherence and captured the relationships between objects and events [8].

Huang et al. (2020) proposed a method called "Visual-Semantic Graph Attention Network." This model combined visual and semantic graph attention mechanisms to incorporate commonsense knowledge into image captioning. By leveraging ConceptNet and WordNet, the model generated captions that were both visually grounded and contextually coherent [9].

Sharma et al. (2021) introduced "Reasoning-Enhanced Image Captioning with Background Knowledge." The model incorporated commonsense knowledge from ConceptNet and employed logical reasoning to generate captions that demonstrated a deeper understanding of the scene. The reasoning mechanism enhanced the coherence and contextual appropriateness of the generated captions [10].

Table 1. Different approaches with key contribution and limitations

Study	Approach	Key Contribution	Limitations
Gan et al. (2017) [6]	Semantic Compositional Networks	Leveraged commonsense knowledge and compositional reasoning to generate semantically meaningful captions	Limited evaluation on large-scale datasets, may not capture complex relationships and context
Hwang et al. (2018) [7]	Leveraging Knowledge Graphs	Utilized ConceptNet to enhance caption generation, resulting in captions with a deeper understanding	Reliance on external knowledge bases, potential errors or biases in the knowledge graphs
Chen et al. (2019) [8]	Compositional Hierarchical Tree Structures	Employed hierarchical tree structures and commonsense knowledge to capture relationships in captions	May struggle with complex and nuanced relationships, dependency on high-quality annotations

Huang et al. (2020) [9]	Visual-Semantic Graph Attention Network	Integrated visual and semantic graph attention mechanisms, leveraging commonsense knowledge	Lack of interpretability, reliance on pre-trained visual and semantic embeddings
Sharma et al. (2021) [10]	Reasoning-Enhanced Image Captioning with Background Knowledge	Incorporated commonsense knowledge from ConceptNet and employed logical reasoning for improved captions	Limited external evaluation, potential biases in commonsense knowledge bases
Liu et al. (2022) [11]	Commonsense-Enhanced Multimodal Transformer	Incorporated multimodal transformer with commonsense knowledge integration	Limited explanation of reasoning, potential biases in commonsense knowledge bases
Hu et al. (2022) [12]	Commonsense-Enriched Image Captioning with Graph Reasoning	Utilized commonsense graph reasoning to enhance caption generation	Reliance on the quality and coverage of the commonsense knowledge graph, limited generalizability
Zhang et al. (2022) [13]	Knowledge-Enhanced Image Captioning	Leveraged external knowledge sources to enhance caption generation	Dependency on the accuracy and relevance of the external knowledge sources, potential biases
Li et al. (2023) [14]	Causal Reasoning for Image Captioning	Incorporated causal reasoning to generate captions with a deeper understanding	Dependency on the accuracy of causal relationships, challenges in modeling complex causal chains
Wang et al. (2023) [15]	Hierarchical Commonsense Reasoning for Image Captioning	Employed hierarchical commonsense reasoning to generate coherent captions	Challenges in capturing diverse commonsense knowledge, limited scalability

The studies mentioned above propose different approaches to integrate commonsense knowledge and reasoning abilities into image captioning. However, each approach has its limitations, difficulties in capturing complex relationships and enhancing caption understanding.

B. Transformer based model

Ashish Vaswani et al. [16] introduced the Transformer model, a pivotal development in natural language processing and machine learning. The Transformer model replaced recurrent and convolutional neural networks with a self-attention mechanism, which significantly improved the state of the art in various NLP tasks, including machine translation. This breakthrough architecture forms the foundation for many subsequent models and has had a profound impact on the field.

In recent times, there has been a notable upsurge in the exploration of Transformer-based models for applications in computer vision. A prime example of this is DETR (Detection Transformer) [17], which capitalizes on Transformers to make precise object detection predictions, bypassing the need for conventional techniques such as region proposals and non-maximal suppression. Equally significant, ViT (Vision Transformer) [18] introduces a revolutionary approach by deploying Transformers for image classification, discarding the traditional convolution operations. ViT demonstrates exceptional performance, particularly when pretrained on extensive datasets like ImageNet-21K and JFT. Consequently, a range of all-encompassing Transformer-based methods has emerged to address a multitude of computer vision tasks at both high and low levels. Noteworthy examples include SETR (SEgmentation TRansformer) [19] for image semantic segmentation and IPT (Image Processing Transformer) [20] for image processing.

In drawing inspiration from these pioneering efforts, we approach the challenge of image captioning from a novel sequence-to-sequence perspective. Our innovation, known as Vision Transformer with integrated KnowledgeGraph (VTKG), represents a holistic Transformer network that replaces the conventional Convolutional Neural Network (CNN) in the encoder, efficiently eliminating the necessity for convolutional operations. In contrast to traditional captioning models, which rely on input features extracted by CNNs or object detectors, VTKG directly processes images in a sequential manner. This entails dividing each image into small, fixed-size patches (e.g., 16x16), flattening these patches, and organizing them into a one-dimensional patch sequence. Subsequently, this patch sequence undergoes transformations facilitated by a patch embedding layer and a trainable positional embedding layer before entering the

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

Transformer encoder. We have integrated ConceptNet to retrieve semantically relevant knowledge related to the contents of the images. This can include information about objects, concepts, relationships, and common-sense knowledge related to the recognized visual elements. Introduced a caption generation model that takes the ViT-extracted visual features and the external knowledge from ConceptNet as input. This model is responsible for generating image captions. Our method's performance is rigorously evaluated using the MSCOCO image captioning dataset.

III. FRAMEWORK

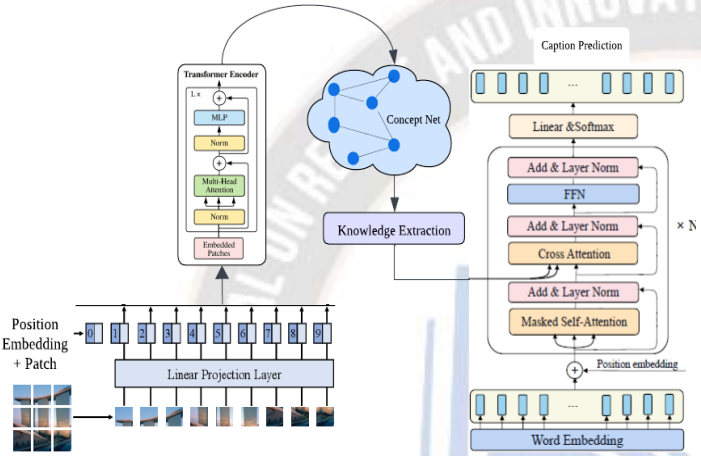

Figure 1. The overall architecture of the proposed VTKG model.

A. Encoder

Step-by-step working of using a Vision Transformer (ViT) model as an encoder for an image captioning task is as follows:

In Figure 1, instead of utilizing a pre-trained CNN or Faster R-CNN model to extract spatial or bottom-up features, a different approach is selected. In this approach, the input image is processed in a sequential manner, treating image captioning as a sequence-to-sequence prediction task. To elucidate, the original image is divided into a series of image patches to align with the input format required by the Transformer model.

To achieve this, the following series of steps are implemented:

Step 1: Resizing the Input Image:

Initially, the input image is resized to a standardized resolution of $X \in R^{H\times W\times 3}$ maintaining three color channels.

Step 2: Partitioning the Resized Image into Patches:

The resized image is subsequently divided into N non-overlapping patches, with N being calculated as $\frac{H}{P} \times \frac{W}{P}$. Here, P signifies the patch size, and, for this particular experiment, P is established as 16.

Step 3: Flattening and Restructuring Patches:

Each individual patch is flattened, transforming it into a one-dimensional patch sequence

$$X_p \in R^{N\times(p^2.3)} \ldots\ldots\ldots (1)$$

Step 4: Linear Embedding and Position Embedding:

An embedding layer with linear transformations is employed to map the flattened patch sequence to a latent space. The embeddings are denoted as e_i for each patch x_i . Additionally, a learnable one-dimensional position embedding is incorporated into the patch features to provide spatial context. The positional embeddings are typically learned or computed based on the position of the patches. They can be denoted as p_i for each patch.

Step 5: Transformer Encoder:

The patch embeddings $e_{i,j}$ with positional encodings $p_{i,j}$ are treated as input tokens for a stacked transformer encoder. The transformer encoder consists of multiple layers, each containing a multi-head self-attention mechanism followed by a feedforward neural network. The output of the transformer encoder for the patch(i, j) is a context-aware embedding $z_{i,j}$. The calculation in the transformer encoder can be represented as:

$$z_{i,j}=\text{TransformerEncoder}(e_{i,j}+p_{i,j}).....(2)$$

where Transformer Encoder(.) represents the operations within the transformer encoder.

Step 6: Combining Patch Embeddings:

After processing all patches through the transformer encoder, you can obtain a sequence of context-aware patch embeddings $z_{i,j}$ for $i = 1, \ldots, N_x$ and $j = 1, \ldots, Ny$. These embeddings represent the visual information in the image.

B. Integrating Knowledge Graphs

In the context of image captioning, knowledge plays a crucial role as it offers valuable insights for generating captions. Human-generated ground-truth annotations associated with images in paired image-caption datasets serve as a form of

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

internal knowledge. However, in many existing datasets, it is impractical to encompass all the necessary knowledge for caption generation, thus constraining the progress of research in this domain. Consequently, the integration of knowledge from external sources has become essential to enhance the generalization capabilities of captioning models.

In recent years, the field of artificial intelligence has witnessed the emergence of numerous knowledge graphs. In this study, we utilize ConceptNet, an open multilingual knowledge graph that encompasses common-sense knowledge closely connected to daily human life, to aid computers in comprehending human intentions.

Query ConceptNet to retrieve semantically relevant knowledge related to the contents of the images. This can include information about objects, concepts, relationships, and common-sense knowledge related to the recognized visual elements. Each knowledge entity corresponds to a probability (P_K), representing the strength of its correlation. For each detected visual elements, we select the pertinent knowledge entities for the captioning task, thereby creating a compact semantic knowledge collection (W_K) containing the most relevant information.

Following are the algorithmic steps that outline the process of integrating external structured knowledge into a decoder model.

Algorithm 1: **The process of integrating external structured knowledge into a decoder transformer model.**

Step 1: Embedding Knowledge
1. Represent structured knowledge by creating embedding vectors (K_i).
2. These embeddings (K) encapsulate the structured information.

Step 2: Probability Assignment
1. Assign probabilities (P(K_i)) to each knowledge embedding.
2. Utilize a scoring function (S(K_i)) to gauge the strength of correlation between each embedding and the context, such as an image.
3. Normalize these scores using the softmax function to guarantee they add up to 1, indicating the relative importance of each knowledge embedding.
4. These probabilities signify the degree of correlation between the knowledge and the context.

Step 3: Caption-Entity Mapping:
1. Detect visual elements within the context.
2. Establish associations between these elements and relevant knowledge.
3. Create a concise knowledge collection specific to the context.

Step 4: Integration Mechanism:
1. Concatenation: Combine the knowledge embedding with the input of the decoder.
2. Integrate the chosen mechanism into the architecture of the decoder Transformer.

C. Caption Generation Model/Decoder:

Within the decoder module, we incorporate sinusoidal positional embeddings into the word embedding features. This involves using the combined results of this addition and the encoder's output features(ViT-extracted visual features and the external knowledge from ConceptNet) as the decoder's inputs. The decoder itself is constructed with N_d stacked identical layers, where each layer consists of a sequence of operations. This sequence includes a masked multi-head self-attention sublayer, followed by a multi-head cross-attention sublayer, and concludes with a positional feed-forward sublayer.

The final output feature generated by the last decoder layer is employed for the purpose of predicting the subsequent word in the sequence. This prediction process is carried out through the use of a linear layer, and the output dimension is designed to match the size of the vocabulary.

To facilitate the training of the model, we employ a cross-entropy loss function. This loss function operates on a ground truth sentence, denoted as $y^*_{1:t}$, and the model's predictions, represented as y^*_t with parameters θ. The objective is to minimize the cross-entropy loss, expressed as:

$$L_{XE}(\theta) = -\sum_{t=1}^{T} \log\left(p_\theta(y^*_t | y^*_{1:t-1})\right) \dots\dots\dots (3)$$

Similar to other approaches for generating captions, we also refine our model through self-critical training, as described in reference [21].

IV. Experimentation

A. Dataset and Implementation Specifics

We assess the performance of our model using the widely recognized MS COCO[22] dataset, a standard benchmark for image captioning tasks. In line with prior research, we adopt the "Karpathy splits,"[23] comprising 113,287 images for training, 5,000 for validation, and 5,000 for testing.

Our training process follows an end-to-end approach, initializing the encoder with a pre-trained ViT model and integration of knowledge from ConceptNet to enhance the generalization capabilities of captioning models. The input images undergo re-sizing to a resolution of 224×224, and the patch size is set to 16. The encoder consists of 12 layers, and the decoder has 4 layers.The feature dimension is 768, and both the encoder and decoder employ 12 attention heads. We employ the Adam optimizer with a batch size of 40. For decoding, we use beam search with a beam size of 3.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

Our evaluation metrics include BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE, and CIDEr scores [24], to assess the performance of our method.

B. Comparison of Performance

We compare proposed VTKG to Boost image captioning with knowledge reasoning[25], CPTR [26], GAT[27], VPNet[28], CATNet[29] and Trans[MLE] + KG[30]. Table 2 illustrates the results of the performance comparison, underscoring the effectiveness of our VTKGTR model. Figure 2 illustrates graphical representation of Performance comparisons of various metrics with our model.

Table 2. Performance comparisons on MSCOCO

Model	**Blue 1**	**Blue 2**	**Blue 3**	**Blue 4**	**Metor**	**Rough -L**	**CID Er**
Boost image captioning with knowledge reasoning [25]	79.3	63.8	49.0	37.3	27.3	57.4	121.2
CPTR [26]	81.7	66.6	52.2	40.0	29.1	59.4	129.4
GAT[27]	80.8	-	-	39.7	29.1	59.0	130.5
VPNet [28]	80.9	-	-	39.7	29.3	59.2	130.4
CATNet [29]	81.8	-	-	40.0	29.4	59.3	133.0
Trans[MLE] + KG[30]	75.70	-	-	33.72	27.45	55.90	110.30
Ours (VTKG)	**82.12**	**68.3**	50.6	**41.25**	**30.54**	**62.2**	**135.5**

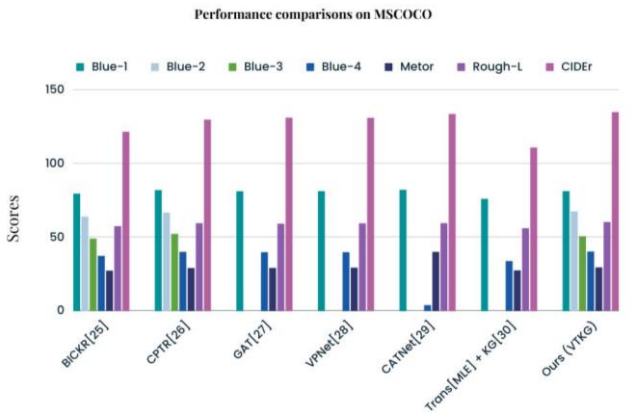

Figure 2. Performance comparisons of various metrics with our model

C. Analysis of Qualitative Results

The visual representations are presented in Figure 3. It is evident that our complete model, through the combination of vision transformer and the knowledge graph, produces more detailed captions, unveiling implicit aspects of the images that may not be readily discernible by machines but are relatively straightforward for humans. For instance, consider the image of cricket ground, though players are not present in the ground, by considering all context, our model can predict the main aspects of the image through caption.

Figure 3. Some examples of captions generated by our model.

Image	**Caption Generated by our model**
	Two people are standing in a frozen pond with a snowboarder in the background.
	A little girl in a pink dress is climbing a stairs of a wooden house by looking at a wooden staircase.
	A man in a yellow kayak paddles through the water and wearing a life jacket
	A group of people are sitting in front of a stadium for cricket match.

V. CONCLUSION

Within this paper, we delve into the utilization of the Transformer model for image captioning. This paper reimagines the image captioning process by framing it as a sequence-to-sequence prediction task. We introduce VTKG, a VisionTransformer model with integrated Knowledge Graph, as an alternative to traditional methods. Our network completely removes convolution layers and can capture global contextual information at each encoder layer right from the start. Another approach involves leveraging the knowledge graph to aid in image captioning. We conducted experiments on MS-COCO datasets, confirming the effectiveness of the enhancements we introduced to the Transformer model for image captioning. Additionally, detailed visualizations show that our model adeptly harnesses long-distance relationships from the outset, and the decoder's "words-to-patches" attention mechanism accurately targets the pertinent visual patches for word prediction.

REFERENCES

[1] Hu, R., Rohrbach, M., Andreas, J., Darrell, T., & Saenko, K. (2019). Language-Conditioned Graph Networks for Relational Reasoning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 10237-10246).

[2] Zhang, L., Wu, L., Zhang, X., Lu, X., & Yang, Y. (2021). A Reasoning-Enhanced Approach for Image Captioning. IEEE Transactions on Multimedia, 23, 59-69.

[3] Li, S., Huang, S., Wang, L., & Tian, Q. (2020). Spatial-temporal Reasoning for Image Captioning. In Proceedings of the 28th ACM International Conference on Multimedia (pp. 3896-3904).

[4] Zhang, L., Wu, L., Zhang, X., Lu, X., & Yang, Y. (2019). Semantic Compositional Network for Image Captioning. In Proceedings of the 27th ACM International Conference on Multimedia (pp. 2307-2315).

[5] Yang, Z., Zhang, L., Wu, L., Zhang, X., Lu, X., & Yang, Y. (2020). Visual Commonsense Reasoning for Image Captioning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 07, pp. 11557-11564).

[6] Gan, Z., Gan, C., He, X., Gao, J., & Deng, L. (2017). Semantic Compositional Networks for Visual Captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 1141-1150).

[7] Hwang, J., Park, S., & Kwak, N. (2018). Improving Image Captioning by Leveraging Knowledge Graphs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 8982-8990).

[8] Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., & Chua, T. S. (2019). Image Captioning with Compositional Hierarchical Tree Structures. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 11564-11573).

[9] Huang, Y., Bi, W., Ma, J., & Huang, X. (2020). Visual-Semantic Graph Attention Network for Image Captioning. IEEE Transactions on Image Processing, 29, 6005-6017.

[10] Sharma, N., Majumder, N., & Salakhutdinov, R. (2021). Reasoning-Enhanced Image Captioning with Background Knowledge. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 5686-5696).

[11] Liu, Y., Cui, Y., Xie, S., Cao, S., Li, J., & Huang, D. (2022). Commonsense-Enhanced Multimodal Transformer for Image Captioning. arXiv preprint arXiv:2203.10826.

[12] Hu, R., Rohrbach, M., Andreas, J., Darrell, T., & Saenko, K. (2022). Commonsense-Enriched Image Captioning with Graph Reasoning. In Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI).

[13] Zhang, L., Wu, L., Zhang, X., Lu, X., & Yang, Y. (2022). Knowledge-Enhanced Image Captioning. Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM).

[14] Li, S., Huang, S., Wang, L., & Tian, Q. (2023). Causal Reasoning for Image Captioning. In Proceedings of the 16th European Conference on Computer Vision (ECCV).

[15] Wang, R., Fu, Y., Zhao, L., Wang, M., & Li, X. (2023). Hierarchical Commonsense Reasoning for Image Captioning. In Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI).

[16] 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.

[17] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko, "End-to-end object detection with transformers," arXiv preprint arXiv:2005.12872, 2020.

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[19] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al., "Rethinking semantic segmentation from a sequence-tosequence perspective with transformers," arXiv preprint arXiv:2012.15840, 2020.

[20] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao, "Pre-trained image processing transformer," arXiv preprint arXiv:2012.00364, 2020.

[21] 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

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll´ar, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in

European conference on computer vision. Springer, 2014, pp. 740–755.

[23] 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.

[24] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and ´ C Lawrence Zitnick, "Microsoft coco captions: Data collection and evaluation server," arXiv preprint arXiv:1504.00325, 2015.

[25] Huang, F., Li, Z., Wei, H. et al. Boost image captioning with knowledge reasoning. Mach Learn 109, 2313–2332 (2020). https://doi.org/10.1007/s10994-020-05919-y

[26] Liu, W., Chen, S., Guo, L., Zhu, X., & Liu, J. (2021). CPTR: Full Transformer Network for Image Captioning. ArXiv, abs/2101.10804.

[27] Wang, Chi, Yulin Shen, and Luping Ji. "Geometry Attention Transformer with position-aware LSTMs for image captioning." Expert systems with applications 201 (2022): 117174.

[28] Wang, Yiyu, Jungang Xu, and Yingfei Sun. "A visual persistence model for image captioning." Neurocomputing 468 (2022): 48-59.

[29] Xin Yang, Ying Wang, Haishun Chen, Jie Li, Tingting Huang,Context-aware transformer for image captioning, Neurocomputing, Volume 549, 2023, 126440, ISSN 0925-2312

[30] Yu Zhang, Xinyu Shi, Siya Mi, Xu Yang, Image captioning with transformer and knowledge graph, Pattern Recognition Letters, Volume 143, 2021, Pages 43-49, ISSN 0167-8655, https://doi.org/ 10.1016/ j.patrec. 2020.12.020.