A Multi-Modal Knowledge Graph for Classical Chinese Poetry

Yuqing Li[1], **Yuxin Zhang**[2], **Bin Wu**[13], **Ji-Rong Wen**[24], **Ruihua Song**[24]*, **Ting Bai**[13]*
[1]Beijing University of Posts and Telecommunications
[2]Renmin University of China
[3]Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia
[4] Beijing Academy of Artificial Intelligence
{liyuqing,wubin,baiting}@bupt.edu.cn
{2020201469,jrwen,rsong}@ruc.edu.cn

Abstract

Classical Chinese poetry has a long history and is a precious cultural heritage of humankind. Displaying the classical Chinese poetry in a visual way, helps to cross cultural barriers in different countries, making it enjoyable for all the people. In this paper, we construct a multi-modal knowledge graph for classical Chinese poetry (PKG), in which the visual information of words in the poetry are incorporated. Then a multi-modal pre-training language model, *i.e.,* PKG-Bert, is proposed to obtain the poetry representation with visual information, which bridges the semantic gap between different modalities. PKG-Bert achieves the state-of-the-art performance on the poetry-image retrieval task, showing the effectiveness of incorporating the multi-modal knowledge. The large-scale multi-modal knowledge graph of classical Chinese poetry will be released to promote the researches in classical Chinese culture area[1].

1 Introduction

Classical Chinese poetry plays an important role in cultural transmission in ancient China, and helps the cultural understanding between the east and the west. Over thousands of years, the language of classical Chinese poetry is greatly different from the modern language, such as the semantic meaning of words and the organization of sentences. For example, the word "wulai" (无赖) in modern Chinese only means rascally, but in ancient Chinese, it also means cute, *e.g.*"I like my cute son most. He is lying in the grass beside the stream, peeling lotus pods" (最喜小儿无赖，溪头卧剥莲蓬), which increases the difficulty to understand the language in classical Chinese poetry. To bridge the culture gap, we propose to display the classical Chinese poetry in an intuitively visual way, which makes it easier to enjoy the beauty of classical Chinese poetry for all the people.

Existing studies in classical Chinese poetry area (Liu et al., 2018; Zhipeng et al., 2019; Wu et al., 2021; Wei et al., 2020; Wang et al., 2021) mainly focus on the generation and analyzing of poetry. Different from them, we make a preliminary attempt to translate the classical Chinese poetry into a multi-modal way to help overcome the cultural barriers among different languages, and make it enjoyable for all the people. To bridge the semantic gap between two modalities (Liu et al., 2019a; Wang et al., 2019), we construct a multi-modal knowledge graph for classical Chinese poetry (**PKG**), in which a wealth of information, such as the allusions and visual information of the words in poetry are incorporated. Specifically, PKG consists of text nodes and image nodes. Each text node is linked to the most similar images by computing the similarity of their representations, which are initialized by pre-training model Bert (Devlin et al., 2019) and BriVL (Huo et al., 2021) respectively, and then fine-tuned in our proposed multi-modal pre-training language model **PKG-Bert**.

With the incorporated image information in PKG, PKG-Bert injects the visual knowledge into the semantic learning of poetry words, making it possible to bridge the semantic gap between different modalities. The effectiveness of our method is verified on the poetry-image retrieval task. We invite five experts in classical Chinese poetry area to screen out the most matching poetry (or image) corresponding to each image (or poetry) query, and then construct a evaluation dataset for classical Chinese poetry-image retrieval task. The contributions of our work are as follows:

- We make a preliminary attempt to construct a

*Co-corresponding authors.

[1] The whole project including codes and data is publicly available at https://github.com/liyuqing1/PKG-Bert.

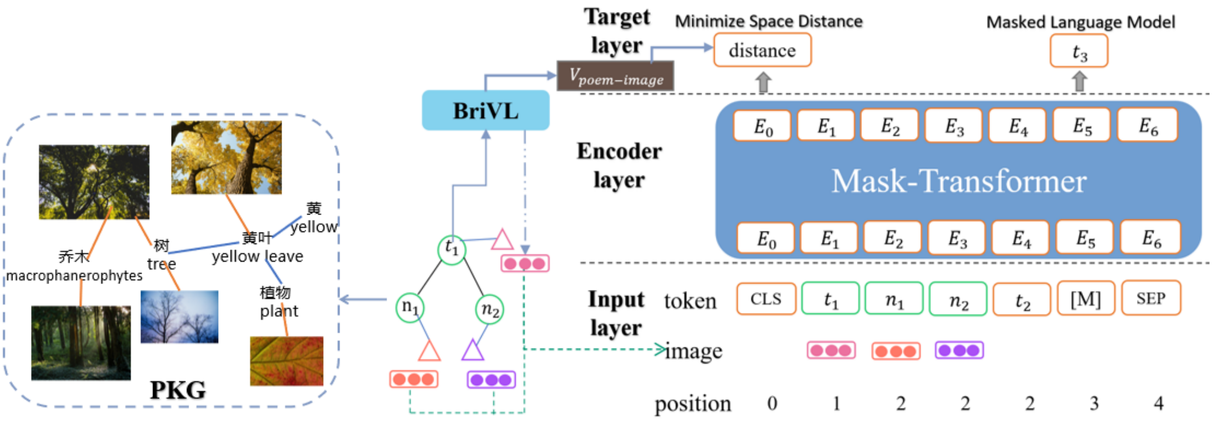


Figure 1: The overview architecture of PKG-Bert. t_1, t_2, t_3 are the words in the input poetry, n_1, n_2 are the text neighbors of t_1 in PKG and triangles represent the image entities.

large-scale multi-modal knowledge graph of classical Chinese poetry (PKG), with 111,514 text nodes and 96,049 image nodes.

- We propose a multi-modal knowledge graph based pre-training language model for classical Chinese poetry (PKG-Bert), making it possible to bridge the gap between the text and visual modalities. Experiments on classical Chinese poetry-image retrieval tasks[2] show the effectiveness of PKG-Bert.

- We contribute a multi-modal knowledge graph PKG in classical Chinese area and a labeled evaluation dataset for poetry-image retrieval task, which will promote the researches on ancient Chinese language understanding.

2 Related Work

In the field of classical Chinese poetry, there are some works on the knowledge graph of poetry (Liu et al., 2020b; Wei et al., 2020; Hong et al., 2020). Liu et al. (Liu et al., 2019b) studies the new word detection in ancient Chinese corpus, and use a semi-automated method to construct a classical Chinese poetry knowledge graph (CCP-KG) (Liu et al., 2020b). They use BERT and GAT to encode the nodes in the text and knowledge graph respectively to classify the sentiment and theme of classical Chinese poetry. KnowPoetry (Hong et al., 2020) semi-automatically constructs a knowledge graph of Tang poetry and poet, and uses inference rules from experts to help the further study of Tang poetry *e.g.,* mining social relationship between poets.

There are also some works on the multi-modal knowledge graph in mordern language (Liu et al., 2019a; Wang et al., 2019). MMKG (Liu et al., 2019a) takes three existing knowledge graphs as the blueprint of the textual part, and uses the search engines to find images highly related with each text node. Richpedia (Wang et al., 2019) obtains texts and images entities from the Wikipedia and search engines, and uses predefined rules to extract relations between text and image entities. However, due to the lack of data, existing studies are hard to apply in the field of classical Chinese poetry. For example, it is difficult for the mainstream search engines to recognize ancient Chinese words, and there is no such large-scale knowledge base like Wikipedia in the field of classical Chinese poetry.

3 Methods

The overall architecture of the multi-modal knowledge graph based pre-training model PKG-Bert is shown in Fig. 1.

3.1 The Construction of PKG

A large-scale multi-modal knowledge graph of classical Chinese poetry, termed as PKG, is constructed by the following steps:

(1) *Poetry data collection*. We crawl the classical Chinese poetry dataset from *Gushiwen*[3] and *Soyun*[4], which contain the content and appreciation of classical Chinese poetry. We collect 905,675 poetry with a range of about three thousand years, that is from the pre-Qin (1000 BC: Before the Common Era) to the Qing dynasty (1800 AD: Anno Domini).

[2] The retrieval system is available at http://facdbe.com:8080/

[3] https://www.gushiwen.cn/

[4] https://sou-yun.cn/

(2) *Text nodes connection.* We keep the words appearing more than 5 times in poetry as the candidate words (145,800 words), then crawl their definitions from *HanDian*[5] and annotations of sememes from *HowNet* (Dong and Dong, 2003; Qi et al., 2019). Sememes are the smallest semantic units in linguistics and well connected in *HowNet*. Based on the connections among sememes, we further add edges between the candidate word and its sememe. As for the word without sememe in its annotation, we use a Sememe Correspondence Pooling (SCorP) model (Du et al., 2020) to predict its sememe, and then build the connections between the word and sememe with positive probability.

(3) *Image nodes connection.* To incorporate the visual information, we add image nodes to PKG. We feed each text node (or its definition in *HanDian*) into the API[6] of a multi-modal pre-training model BriVL (Huo et al., 2021), which returns the top 10 similar images of the input text. Then we make connections between the text node and the highly related image nodes.

Totally, the multi-modal knowledge graph PKG contains **111,514 text nodes** (including the words in poetry and sememes in *HowNet*), **96,049 image nodes**, and **1,369,632 edges**.

3.2 PKG-Bert

To learn the representation of poetry, we propose a multi-modal pre-training language model for classical Chinese poetry, termed as PKG-Bert. The architecture of PKG-Bert is shown in Fig. 1, which is divided into three parts: input layer, encode layer and target layer.

(1) *Input layer.* Input layer consists of three types of information: token, image entity, and position. **Tokens** include the words in poetry and linked text entities in PKG. As shown in Fig. 1, given a poetry $\mathcal{T} = \{t_1, t_2, ..., t_n\}$, where t_i represents a word in the poetry, we first insert special tokens, *i.e.,* [CLS] and [SEP], to tag the start and end of the poetry. Take the word t_1 for example, we then insert its neighbors (*i.e.,* n_1 and n_2) in PKG behind it. The embedding of each token is defined as $\mathbf{v}_{token} \in \mathcal{R}^H$, where H is the hidden dimension. **Image entity** refers to the image node in PKG that linked to the tokens. Given a token, we encode its linked image entities by BriVL (Huo et al., 2021), and then use average pooling and linear transformation operations to obtain its visual representation $\mathbf{v}_{image} \in \mathcal{R}^H$. **Position** records the order of words in a poetry. Note that the inserted text entities in PKG and the next word in poetry are encoded with the same position. For example, n_1, n_2 and t_2 are encoded with the same position embedding $\mathbf{v}_{pos} \in \mathcal{R}^H$. Finally, the input representation of a token is defined as:

$$\mathbf{v}_{\text{Input}} = \mathbf{v}_{\text{token}} + \mathbf{v}_{\text{image}} + \mathbf{v}_{\text{pos}}. \tag{1}$$

(2) *Encode layer.* The encode layer uses Mask-Transformer blocks (Liu et al., 2020a) to extract features from the input. Given the input representation $\mathbf{v}_{\text{Input}}$, the out representation of the ith transformer block $\mathbf{v}^i$ ($\mathbf{v}^0 = \mathbf{v}_{\text{Input}}$) is defined as:

$$\begin{aligned} Q^i &= \mathbf{v}^{i-1}W_q, \\ K^i &= \mathbf{v}^{i-1}W_k, \\ V^i &= \mathbf{v}^{i-1}W_v, \\ \text{Atten}^i &= \text{softmax}\left(\frac{Q^i K^{i^\top} + \text{VM}^i}{\sqrt{d_k}}\right), \\ \mathbf{v}^i &= \text{Atten}^i V^i, \end{aligned} \tag{2}$$

where Q, K, V are the query, key and value vectors of attention mechanism, W_q, W_k, W_v are trainable parameters, d_k is the dimension of Q and K, and VM is the parameter in self-attention block to prevent the inserted entities from disturbing the inference of the words.

(3) *Target layer.* We use Masked Language Model (MLM) (Devlin et al., 2019) to train the language model. The words are randomly masked with a probability of 15% in three mask methods, *i.e.,* replaced by [MASK], replaced by random word, and unchanged with the probabilities of 80%, 10% and 10% respectively.

To overcome the semantic gap between textual and visual modalities, we minimize the space distance between poetry representations learned in PKG-Bert and BriVL (Huo et al., 2021) to keep the poetry and its visual information in the same learning space. It is optimized by:

$$Loss_{\text{space}} = 1 - \cos\left(\mathbf{V}_{\text{PKG-Bert}}, \mathbf{V}_{\text{BriVL}}\right), \tag{3}$$

where $\mathbf{V}_{\text{PKG-Bert}}$ and $\mathbf{V}_{\text{BriVL}}$ are the vectors of a poetry in PKG-Bert and BriVL respectively.

The final optimization function is defined as:

$$Loss = Loss_{\text{MLM}} + Loss_{\text{space}}. \tag{4}$$

[5] https://www.zdic.net/

[6] http://buling.wudaoai.cn/pc/htmls/input.html

Table 1: The performance of different methods.

Model	Poetry-to-Image Task			Image-to-Poetry Task		
	HR@5	NDCG@5	MRR	HR@5	NDCG@5	MRR
BriVL	0.26	0.4918	0.2227	0.56	0.7228	0.3924
CLIP	0.22	0.5192	0.1826	0.42	0.5645	0.2153
ViLT	0.24	0.4440	0.1667	0.62	0.6239	0.4509
PKG-Bert	**0.46**	**0.7011**	**0.3607**	**0.72**	**0.7350**	**0.5382**
-Visual	0.40	0.6802	0.3312	0.64	0.6691	0.4213
-PKG	0.16	0.3499	0.1120	0.56	0.6143	0.3462
Improv	**77%** ↑	**35%** ↑	**62%** ↑	**16%** ↑	**2%** ↑	**19%** ↑

Table 2: The statistic of dataset.

# Poetry	905,675
# Images	200,000
# Text entities in PKG	111,514
# Image entities in PKG	96,049
# Edges in PKG	1,369,632

4 Experiments

We make a preliminary attempt to design a poetry-image retrieval task, which could be well used to evaluate the effectiveness of PKG-Bert.

4.1 Experimental Settings

We collect the classical Chinese poetry from *Gushiwen* and *Souyun* and images from *Unplash*[7] which is a free community of photographers and provides a large number of high-definition images. The statistics of the dataset is shown in Table 2.

To evaluate the model performance, we randomly select 50 images (or poetry) as the queries, choose the top 10 poetry (or image) candidates (selected form 30,000 poetry and 200,000 images) from each model, and then take them together to construct the evaluation set. We invite five experts in classical Chinese area to label the similarity scores of the candidates for each query, which is mainly decided by the number of related objects in the candidates. We take the average score from five experts as the final score of a candidate.

4.1.1 Baseline Methods and Metrics

We compare PKG-Bert with recent multi-modal pre-training models, including:

- BriVL (Huo et al., 2021): the largest Chinese multi-modal pre-training model, and pre-trained with 30 million text-image pairs.
- CLIP (Radford et al., 2021): a widely used multi-modal pre-training model with a larger scale of training data (400 million pairs).
- ViLT (Kim et al., 2021): a multi-modal pre-training model with a rigorous inter-modal interaction scheme, and requires strong semantic correlation between the input text-image pairs (9 million pairs).

For fair comparison, we translate classical Chinese poetry into English before feeding into CLIP and ViLT. All the models above are evaluated on two tasks: poetry-to-image and image-to-poetry retrieval. We adopt the widely used metrics: HR@K (Hit Ratio), NDCG (Normalized Discounted Cumulative Gain) (Järvelin and Kekäläinen, 2002) and MRR (Mean Reciprocal Rank) (Voorhees et al., 1999) to evaluate the model performance. The details of implementation are introduced in appendix A.1.

4.2 Main Results

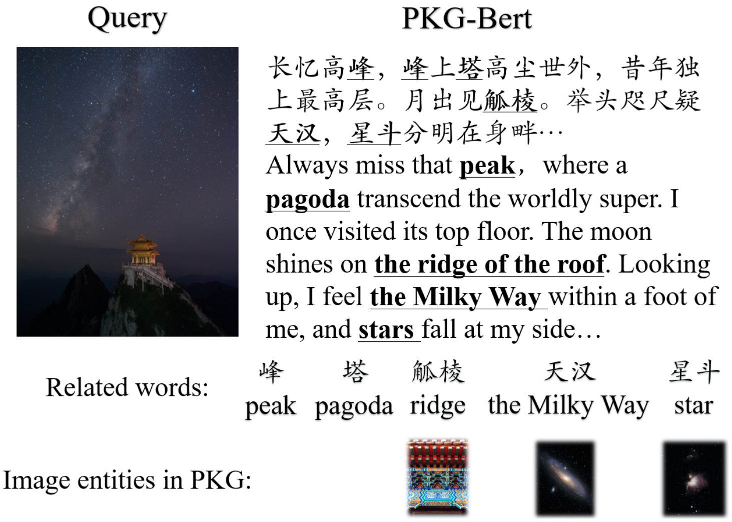
The performance of different models are shown in Table 1, we can see that:

(1) PKG-Bert achieves the best performance in two tasks. Although without the pre-training process of text-image pairs, PKG-Bert still performs the best, indicating the usefulness of incorporating the multi-modal knowledge graph PKG in poetry-image retrieval task.

(2) PKG-Bert performs better than its variants, *i.e.,* -Visual and -PKG, which learns without the visual information in PKG and without the entire PKG respectively. This indicates that both the visual information and the incorporated extra knowledge in KG help to overcome the semantic gap between two modalities.

(3) PKG-Bert obtains a greater improvement on poetry-to-image task (with an average improve-

[7] https://unsplash.com/



长忆高峰，峰上塔高尘世外，昔年独上最高层。月出见觚棱。举头咫尺疑天汉，星斗分明在身畔···

Always miss that **peak**，where a **pagoda** transcend the worldly super. I once visited its top floor. The moon shines on **the ridge of the roof**. Looking up, I feel **the Milky Way** within a foot of me, and **stars** fall at my side…

Figure 2: An example of image-to-poetry. Underlined words are highly related objects in the candidate poetry.

ment of 57%) than image-to-poetry task (12%). This may lie in that the understanding of query is more important in retrieval tasks. In the poetry-to-image task, the semantic of the poetry can be much more well learned in PKG-Bert compared with other multi-modal language models.

4.3 Case Study

We present a case (see in Fig. 2) of the image-to-poetry retrieval task. For an image query of "the starry sky (星空)", five highly related objects, *i.e.,* "peak (峰)", "pagoda (塔)", "ridge (觚棱)", "the Milky Way (天汉)" and "star (星斗)" are retrieved in the candidate poetry in PKG-Bert. While in ViLT, only three objects, *i.e.,* "tower (塔)", "mountain (山)", "night (宵)" are retrieved.

It is difficult to retrieve the related classical Chinese poetry with "the starry sky (星空)" objects, since that most of the multi-modal pre-training language models are based on modern language, while the words describing "the starry sky " in classical Chinese poetry (*e.g.,* "天汉", "星斗") rarely appear in modern Chinese, which increases the difficulty to retrieve the related classical Chinese poetry. In PKG-Bert, the words in classical Chinese poetry are linked to the sememes, which are associated with the objects in modern language, helping bridges the gap between modern and ancient Chinese language. Besides, the linked visual information of poetry words also helps the retrieval task. More details are shown in appendix A.2.

5 Conclusions

We construct a multi-modal knowledge graph for classical Chinese poetry (PKG), and propose a multi-modal pre-training language model PKG-Bert. It incorporates the visual information into the semantic learning of classical Chinese poetry, which helps to promote the researches in classical Chinese language understanding. The poetry-image retrieval application also helps to cross the cultural barriers in different countries, making it enjoyable for all the people.

Limitations

This paper uses multi-modal information to enhance the pre-training language model in classical Chinese poetry area, which could be extended to the general Chinese poetry in the future. In addition, the objects in a poetry may not exactly match the images in candidate image set. Thus, collecting a larger scale image set will further improve the model performance.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant No.62102038, Beijing Academy of Artificial Intelligence (BAAI), the National Natural Science Foundation of China under Grant No.61972047, 62236003 and the NSFC-General Technology Basic Research Joint Funds under Grant U1936220.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Zhendong Dong and Qiang Dong. 2003. Hownet-a hybrid language and knowledge resource. In *International Conference on Natural Language Processing and Knowledge Engineering*, pages 820–824. IEEE.

Jiaju Du, Fanchao Qi, Maosong Sun, and Zhiyuan Liu. 2020. Lexical sememe prediction using dictionary definitions by capturing local semantic correspondence. *Journal of Chinese Information Processing*.

Dan Hendrycks and Kevin Gimpel. 2017. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. In *International Conference on Learning Representations*.

Liang Hong, Wenjun Hou, and Lina Zhou. 2020. Knowpoetry: A knowledge service platform for tang poetry research based on domain-specific knowledge graph. *Library Trends*, 69(1):101–124.

Yuqi Huo, Manli Zhang, Guangzhen Liu, Haoyu Lu, Yizhao Gao, Guoxing Yang, Jingyuan Wen, Heng Zhang, Baogui Xu, Weihao Zheng, et al. 2021. Wenlan: Bridging vision and language by large-scale multi-modal pre-training. *arXiv preprint arXiv:2103.06561*.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. *The Thirty-eighth International Conference on Machine Learning*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Dayiheng Liu, Quan Guo, Wubo Li, and Jiancheng Lv. 2018. A multi-modal chinese poetry generation model. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020a. K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2901–2908.

Ye Liu, Hui Li, Alberto Garcia-Duran, Mathias Niepert, Daniel Onoro-Rubio, and David S Rosenblum. 2019a. Mmkg: multi-modal knowledge graphs. In *European Semantic Web Conference*, pages 459–474. Springer.

Yutong Liu, Bin Wu, and Ting Bai. 2020b. The construction and analysis of classical chinese poetry knowledge graph. *Journal of Computer Research and Development*, 57(6):1252.

Yutong Liu, Bin Wu, Tao Xie, and Bai Wang. 2019b. New word detection in ancient chinese corpus. *Journal of Chinese Information Processing*, 33(1):46–55.

Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Qiang Dong, Maosong Sun, and Zhendong Dong. 2019. Openhownet: An open sememe-based lexical knowledge base. *arXiv preprint arXiv:1901.09957*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Ellen M Voorhees et al. 1999. The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82.

Meng Wang, Guilin Qi, HaoFen Wang, and Qiushuo Zheng. 2019. Richpedia: A comprehensive multimodal knowledge graph. In *Joint International Semantic Technology Conference*, pages 130–145. Springer.

Qing Wang, Xiumei Wang, Weiping Liu, and Guannan Chen. 2021. Predicting the chinese poetry prosodic based on a developed bert model. In *2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, pages 583–586.

Yuting Wei, Huazheng Wang, Jiaqi Zhao, Yutong Liu, Yun Zhang, and Bin Wu. 2020. Gelaigelai: A visual platform for analysis of classical chinese poetry based on knowledge graph. In *2020 IEEE International Conference on Knowledge Graph (ICKG)*, pages 513–520.

Chunlei Wu, Jiangnan Wang, Shaozu Yuan, Leiquan Wang, and Weishan Zhang. 2021. Generate classical chinese poems with theme-style from images. *Pattern Recognition Letters*.

Guo Zhipeng, Xiaoyuan Yi, Maosong Sun, Wenhao Li, Cheng Yang, Jiannan Liang, Huimin Chen, Yuhui Zhang, and Ruoyu Li. 2019. Jiuge: A human-machine collaborative chinese classical poetry generation system. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 25–30.

A Appendix

A.1 Implementation Details

PKG-Bert is initialized by Bert (Devlin et al., 2019), and a grid search is applied to find the optimal settings. We use the Adam (Kingma and Ba, 2015) as the optimizer with a weight decay rate of 0.01 and the learning rate is 2e-5 (selected among 2e-6, 2e-5, and 2e-4). The number of transformer blocks in PKG-Bert is 12, the number of heads in transformer is 12, the dimension of the vectors in transformer (d_k) is 64, the activation function is GELU (Hendrycks and Gimpel, 2017), the hidden dimension of embeddings (H) is 768, and the batch size is set to 16. The number of total parameters is 122M. Our model is implemented by Pytorch 1.6.0 with NVIDIA GeForce RTX 2080, and is trained for 84 epochs in which the scales of train and validation sets are 200,000 and 10,000 respectively, and the inference time is 21 milliseconds.

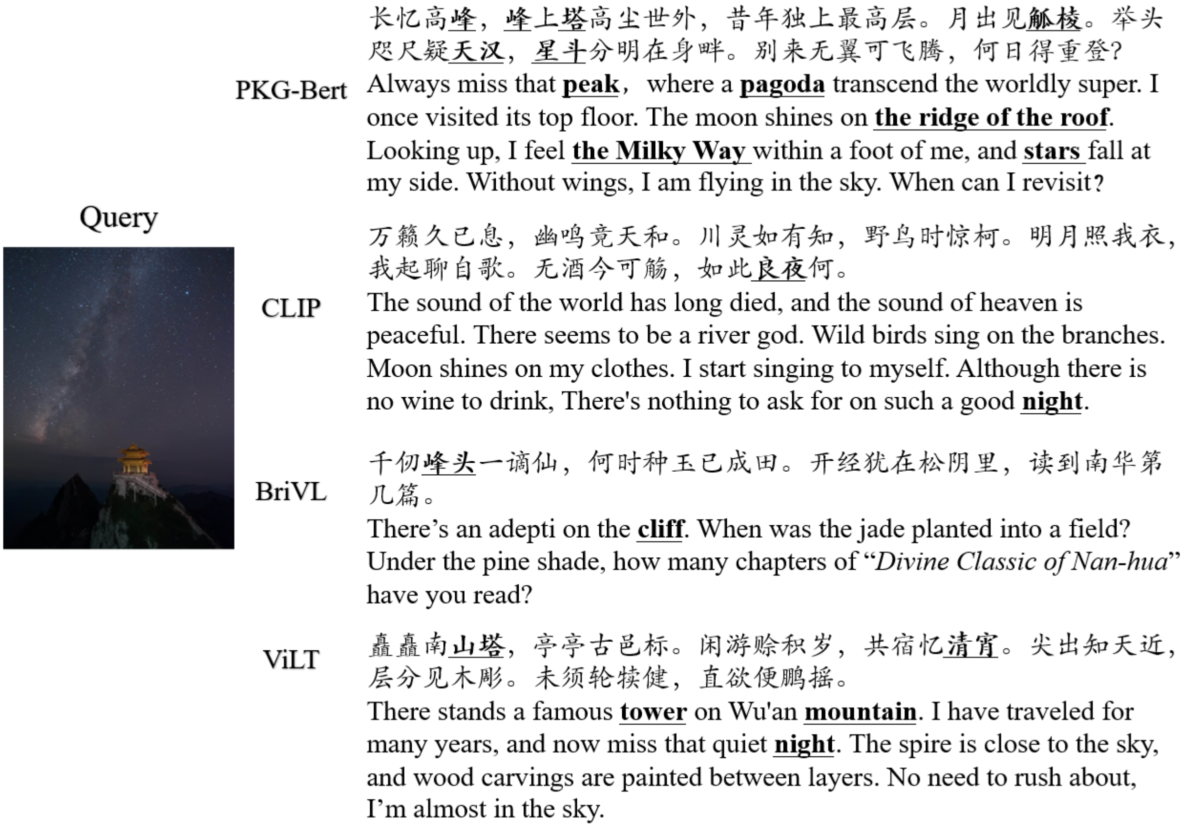
A.2 Case Studies

A.2.1 Image-to-Poetry Task

Fig. 3 and Fig. 4 show the retrieved objects in PKG-Bert, CLIP, BriVL and ViLT, *i.e.,* five objects “peak (峰)”, “pagoda (塔)”, “the ridge of the roof (觚棱)”, “the Milky Way (天汉)” and “stars (星斗)” in PKG-Bert; only “night (夜)” in CLIP; only “cliff (峰头)” in BriVL; three objects “tower (塔)”, “mountain (山)”, “night (宵)” in ViLT, which shows that PKG-Bert can achieve better performance on the image-to-poetry retrieval task.

A.2.2 Poetry-to-Image Task

Fig. 5 and Fig. 6 show the retrieved objects in PKG-Bert, CLIP, BriVL and ViLT, *i.e.,* three objects “cloud (云雾)”, “river (江)”, “tree (树)” in PKG-Bert; only “tree (树)” in BriVL; two objects “people (人)” and “tree (树)” in CLIP; only “people (人)” in ViLT, indicating the power of PKG-Bert in the poetry-to-image retrieval task.



Query

PKG-Bert

长忆高峰，峰上塔高尘世外，昔年独上最高层。月出见觚棱。举头咫尺疑天汉，星斗分明在身畔。别来无翼可飞腾，何日得重登?

Always miss that **peak**， where a **pagoda** transcend the worldly super. I once visited its top floor. The moon shines on **the ridge of the roof**. Looking up, I feel **the Milky Way** within a foot of me, and **stars** fall at my side. Without wings, I am flying in the sky. When can I revisit?

CLIP

万籁久已息，幽鸣竟天和。川灵如有知，野鸟时惊柯。明月照我衣，我起聊自歌。无酒今可觞，如此良夜何。

The sound of the world has long died, and the sound of heaven is peaceful. There seems to be a river god. Wild birds sing on the branches. Moon shines on my clothes. I start singing to myself. Although there is no wine to drink, There's nothing to ask for on such a good **night**.

BriVL

千仞峰头一谪仙，何时种玉已成田。开经犹在松阴里，读到南华第几篇。

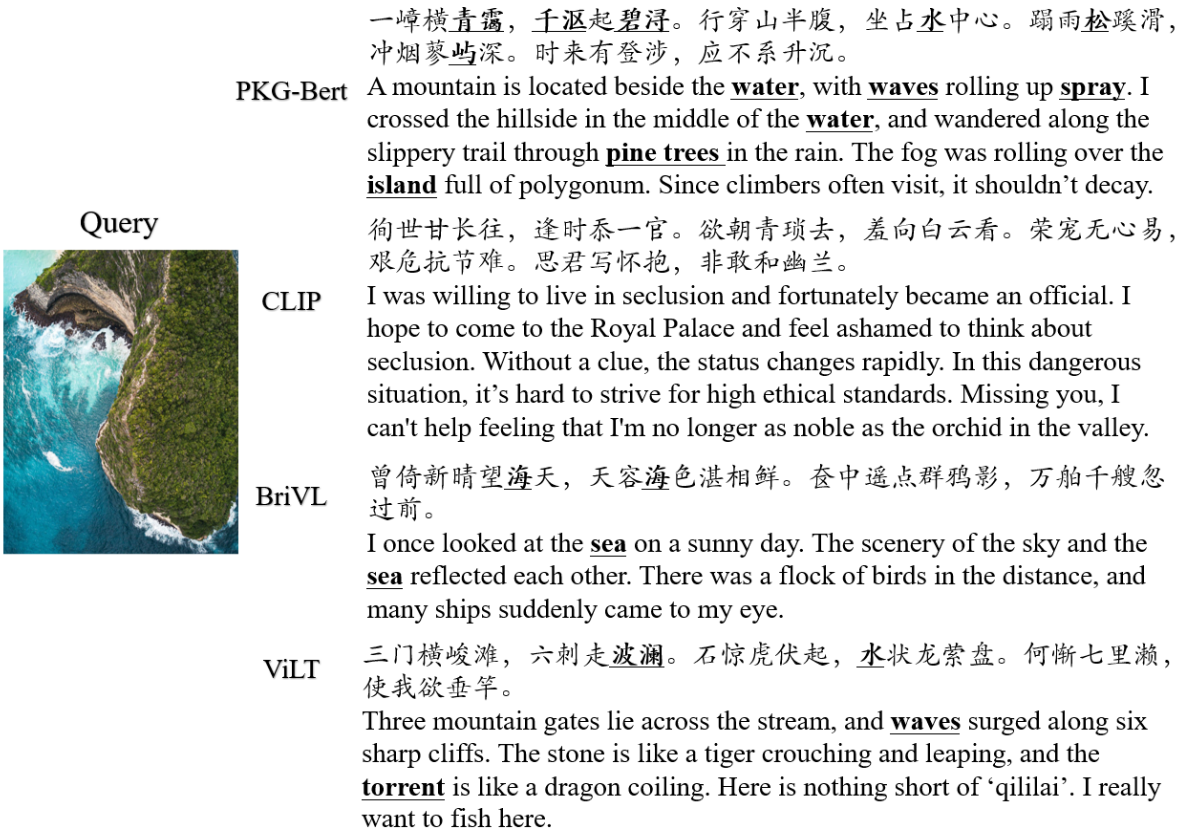
There's an adepti on the **cliff**. When was the jade planted into a field? Under the pine shade, how many chapters of "*Divine Classic of Nan-hua*" have you read?

ViLT

矗矗南山塔，亭亭古邑标。闲游赊积岁，共宿忆清宵。尖出知天近，层分见木彫。未须轮犊健，直欲便鹏摇。

There stands a famous **tower** on Wu'an **mountain**. I have traveled for many years, and now miss that quiet **night**. The spire is close to the sky, and wood carvings are painted between layers. No need to rush about, I'm almost in the sky.

Figure 3: The retrieved poetry for the query of a photo with the starry sky. Underlined words are related objects in candidate poetry.



Query

PKG-Bert

一嶂横青霭，千沤起碧浔。行穿山半腹，坐占水中心。蹋雨松蹊滑，冲烟蓼屿深。时来有登涉，应不系升沉。

A mountain is located beside the **water**, with **waves** rolling up **spray**. I crossed the hillside in the middle of the **water**, and wandered along the slippery trail through **pine trees** in the rain. The fog was rolling over the **island** full of polygonum. Since climbers often visit, it shouldn't decay.

CLIP

徇世甘长往，逢时忝一官。欲朝青琐去，羞向白云看。荣宠无心易，艰危抗节难。思君写怀抱，非敢和幽兰。

I was willing to live in seclusion and fortunately became an official. I hope to come to the Royal Palace and feel ashamed to think about seclusion. Without a clue, the status changes rapidly. In this dangerous situation, it's hard to strive for high ethical standards. Missing you, I can't help feeling that I'm no longer as noble as the orchid in the valley.

BriVL

曾倚新晴望海天，天容海色湛相鲜。奁中遥点群鸦影，万舶千艘忽过前。

I once looked at the **sea** on a sunny day. The scenery of the sky and the **sea** reflected each other. There was a flock of birds in the distance, and many ships suddenly came to my eye.

ViLT

三门横峻滩，六刺走波澜。石惊虎伏起，水状龙萦盘。何惭七里濑，使我欲垂竿。

Three mountain gates lie across the stream, and **waves** surged along six sharp cliffs. The stone is like a tiger crouching and leaping, and the **torrent** is like a dragon coiling. Here is nothing short of 'qililai'. I really want to fish here.

Figure 4: The retrieved poetry for the query of a photo with an island covered with trees. Underlined words are related objects in candidate poetry.

独有宦游人，偏惊物候新。云霞出海曙，梅柳渡江春。
Only those **officials** assigned for away-from-home posts
are particularly conscious of nature's changes in seasons.
Morning glow of **sun** risen from **sea** bounces off **clouds**.
Plum and willow **trees** across the **river** bring along spring.

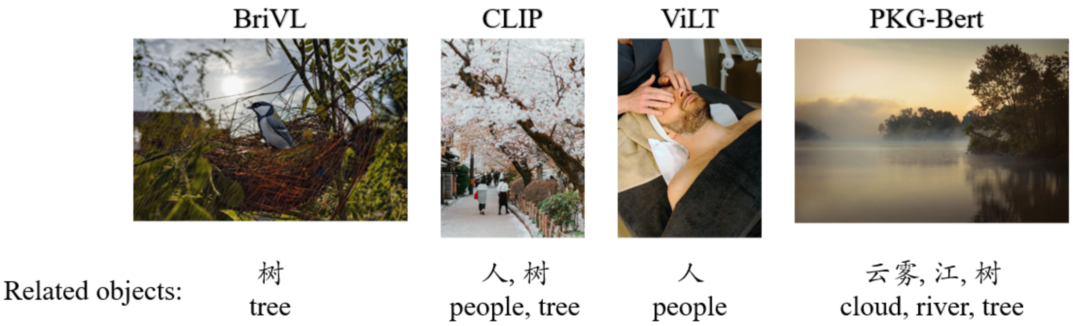


Figure 5: The retrieved images for the query of a poetry of a lake in fog. Underlined words are related objects in candidate images.

上弦如半璧，初魄似蛾眉。渡云光忽驶，中天影更迟。
The moon at the first quarter is like half a white jade, and **the new
moon** is like a woman's eyebrows. The **moon** moved quickly through the
clouds, but I could not feel its movement during the **upper culmination**.

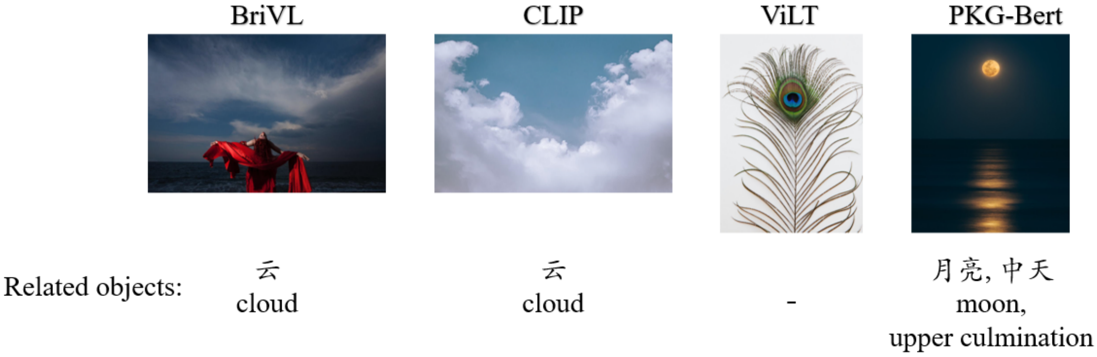


Figure 6: The retrieved images in query of a poetry of the moon. Underlined words are related objects in candidate images.