
Neural Poetry Generation with Visual Inspiration

Zhaoyang Li

Simon Fraser University
zhaoyang_li@sfu.ca

Ge Shi

Simon Fraser University
shiges@sfu.ca

Haipeng Li

Simon Fraser University
haipengl@sfu.ca

Xi Yang

Simon Fraser University
xya81@sfu.ca

Abstract

We present a model that automatically generates poems from images. Our approach leverages datasets of correlated images and poems to learn about the correspondences between visual and textual features as well as datasets of poems only to learn about the poetry language model. Our feature extraction module utilizes convolutional neural networks, BERT, and an alignment model to obtain an aligned, multimodal embedding of images and poems. We describe a recurrent neural network generator that uses the aligned features to produce poetic texts. We then describe an adversarial training process with a discriminator to improve the performance of the generator. We demonstrate that our model is able to generate poems that depict input images and read like human-written.

1 Introduction

Modern deep learning models enabled effective visual information extraction by convolutional neural networks in computer vision [6]. In the field of natural language processing, language models derived from recurrent neural networks are able to generate natural language sentences from intermediate representations that encode information [16]. In the real world, humans can make descriptions from what they see. This process involves the understanding of the image and the description of the understanding with specific sentences, which can be regarded as a combination of extracting visual information and generating natural language texts. A series of works [9, 17, 18] successfully built models to generate captions from images, demonstrating it approachable to generate natural language texts from images.

Mostly, image captions are structured and ordered language. Compared to them, some other types of natural languages are less structured and ordered, such as poetry. Generating these unstructured languages from images is more challenging since there are less visible patterns in them. In addition to the uncertainty, generating poem requires searching for poetic symbols from image elements rather than simply summarizing the visual information. Inspired by [13], our project targets at generating poems from images.

The contributions of this project are: (1) We proposed a joint training method to learn a common embedding space that matches the information from parallel image-poem pairs by aligning image features with textual features. (2) We constructed a recurrent neural network decoder with long short-term memory units to generate poems from extracted image features. (3) We introduced an adversarial discriminator to improve the quality of generated poems.

2 Related Work

2.1 Poetry Generation

Poetry generation is a long-studied problem in natural language processing. Since poetry languages usually contain special rhythms, one way to form poems is to generate rhythmic language. [7] studied generating poems according to rhythmic rules. [5] demonstrated that generating poems has a certain similarity with machine translation, thus reasonable to transfer methods between these tasks. In recent years, recurrent neural networks with long short-term memory units enabled building effective language models to generate natural language, and a lot of research studied poem generation with recurrent neural networks. [20] was the first work trying to generate Chinese poems using plain RNNs. [4] integrated RNN with a finite state acceptor to maintain the rhythmic features at generating. [12] introduced an adversarially trained discriminator to enhance the quality of generated poems, of which the method is similar to ours.

2.2 Image Description Generation

Prior to the flourishing of neural networks, image captioning research mainly targeted at bidirectional image and description retrieval, or filling templates with image information. With CNNs and RNNs, powerful models can be built to generate captions from images. [17, 9] were successful attempts for building LSTM decoder to generate image caption. These models are further improved by techniques such as attention mechanism. Recent researches also introduced adversarial discriminator [19] or adversarial training method [1] to the task and achieved further enhancement.

3 Approach

The goal of our model is to generate poems from images that have a strong semantic correlation. This process involves extracting image features with convolutional neural networks (Figure 1a) and generating poetry texts with a recurrent neural network (Figure 1b). For the shortage of parallel image-poem pairs, we utilize another language representation model to embed textual features and an alignment model to project image features and textual features alike to a common embedding space. We use a discriminator to improve the performance of the generator.

3.1 Image Feature Extraction

The Image Feature Extraction module takes an image as input and produces a vector as its feature representation. This process is modeled with convolutional neural networks. Following a prior work [13], we believe that object-wise, scenic, and sentimental factors in images are critical for poetry. Therefore we use three convolutional neural networks to find the respective features in the input image. The backbones of these CNNs are a commonly used ResNet-50 [6]. The Object CNN is pre-trained on ImageNet; the Scene CNN is pre-trained on a scene recognition dataset [21]; and the Sentiment CNN is pre-trained on ImageNet and fine-tuned on a visual sentiment analysis dataset [2]. Outputs of the penultimate layers (each with dimension 2048) of the CNNs are concatenated together to represent the 6144-dimensional image feature v .

$$v_{Obj} = \text{CNN}_{Obj}(I) \quad (1)$$

$$v_{Sce} = \text{CNN}_{Sce}(I) \quad (2)$$

$$v_{Sen} = \text{CNN}_{Sen}(I) \quad (3)$$

$$v = [v_{Obj}; v_{Sce}; v_{Sen}] \quad (4)$$

3.2 Textual Feature Encoding and Feature Alignment

The ideal method would be to train our model on parallel image-poem pairs. However, the matching of images and poems requires great manual labor and existing datasets are relatively small-scale [13]. To tackle the shortage of parallel dataset, we bootstrap the training by introducing a common embedding space for both images and poems. We use BERT [3] to encode poem sentences into

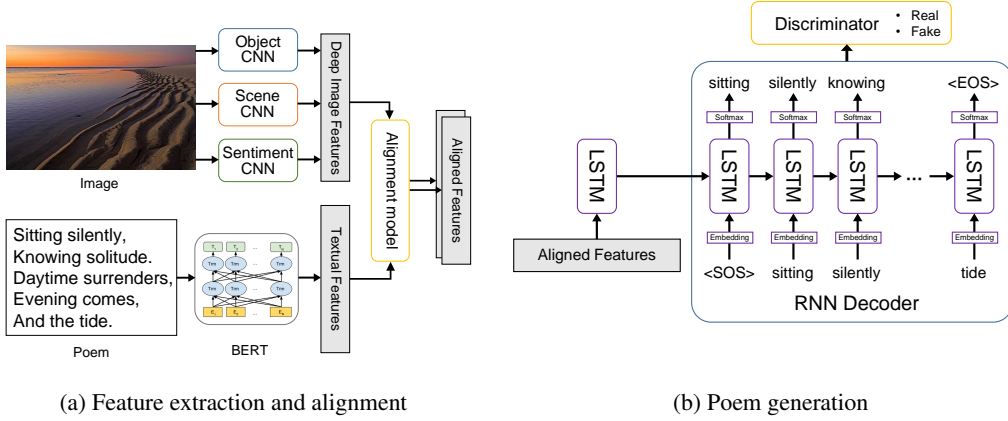


Figure 1: Illustration of our model

a 768-dimensional vector representation t . We project image feature v and textual feature t to a d -dimensional embedding space:

$$t = \text{BERT}(P) \quad (5)$$

$$h_v = W_v v + b_v \quad (6)$$

$$h_t = W_t t + b_t \quad (7)$$

where $W_v \in \mathbb{R}^{d \times 6144}$, $b_v \in \mathbb{R}^d$, $W_t \in \mathbb{R}^{d \times 768}$, and $b_t \in \mathbb{R}^d$ are learnable parameters; h_v and h_t are d -dimensional embedding of images and poems, respectively. We then interpret the dot product $h_v^T h_t$ in this embedding space as the similarity between image embedding h_v and poem embedding h_t . The parameters are learned by minimizing the ranking loss \mathcal{L}_r [10]:

$$\mathcal{L}_r = \sum_{h_v} \sum_{h'_t} \max(0, \alpha - h_v^T h_t + h_v^T h'_t) + \sum_{h_t} \sum_{h'_v} \max(0, \alpha - h_t^T h_v + h_t^T h'_v) \quad (8)$$

where h_v and h_t are paired embedding, h'_t ranges over textual embedding unpaired with h_v , h'_v ranges over image embedding unpaired with h_t , and α is a margin constant. With this objective minimized, we consider the paired embedding h_v and h_t to be aligned as a multimodal vector representation for images and poems alike, and can be used equivalently in the subsequent networks. Given this alignment, we are then able to pre-train our model with textual features on a poem-only dataset and fine-tune our model with image features on a small-scale parallel dataset.

3.3 Poem Generation

The Poem Generation module takes an aligned embedding feature as input and outputs a series of words to form a poetic text that reflects the feature. We apply a recurrent neural network decoder to generate the target poem sentences. We use long short-term memory units in this decoder. The aligned feature is fed into a LSTM cell as an activation. We use the hidden state of the LSTM cell to initialize the hidden state of the LSTM decoder. At each time-step, the LSTM decoder takes a word embedding w_i as input, produces a distribution p_i over the vocabulary, and sample from this distribution to obtain the output word y_i . We apply weight tying [8, 15] in the decoder, sharing the word embedding matrix W_e and the linear transformation from the hidden state to output. We employ teacher forcing and adversarial training at training and beam search at inference. Dropout is applied to avoid overfitting.

$$p_i = \text{softmax}(W_e \cdot \text{RNN}(w_i)) \quad (9)$$

$$y_i \sim p_i \quad (10)$$

3.4 Adversarial Training

Besides teacher forcing, we introduce a discriminator to improve the performance of the decoder during training. The discriminator is a binary classifier that takes a poem as input and produces probabilities whether the input poem is real (i.e. from the dataset) or fake (i.e. generated by the decoder). We use the same word embedding matrix with the decoder and a bidirectional LSTM to encode the input. We perform logistic regression on the last hidden state of the BiLSTM to produce the output.

We regard the decoder as an agent. The agent performs an action of generating poem \mathbf{y} and gets rewarded according to the action. The reward r is decided by the discriminator and is defined to be the probability of the generated poem being real. We maximize the objective function $J(\theta)$:

$$J(\theta) = \sum_{\mathbf{y} \in \mathbb{Y}} p_{\theta}(\mathbf{y}) r(\mathbf{y}) = E_{\mathbf{y} \sim p_{\theta}} r(\mathbf{y}) \quad (11)$$

where \mathbb{Y} is the space of all possible generated poems and $p_{\theta}(\mathbf{y})$ is the probability of generating poem \mathbf{y} parameterized by policy θ . In practice we use Monte-Carlo sampling to approximate the expectation and the parameters can be updated with policy gradient methods. During training, the policy gradient is combined with the gradient from teacher forcing.

4 Experiments

4.1 Training Details

We use two poem datasets (MultiM-Poem and UniM-Poem) collected by [13] in order to facilitate this project. MultiM-Poem dataset is composed of 8,292 image-poem pairs, and UniM-Poem is a larger dataset with 93,265 poems. Some examples from both datasets can be seen in Figure 2.

In the alignment model, the embedding space has dimension $d = 512$. The margin constant α is set to be 0.2. In the RNN decoder model, we use 256-dimensional word embedding and LSTM with 256-dimensional hidden state. Dropout rate is 0.2.

Training strategy

1. Train the Sentiment CNN on Image Sentiment Polarity [2]. The Object CNN and Scene CNN we use are pre-trained [21].
2. Fix the parameters in the CNNs and BERT; train the alignment model on MultiM-Poem.
3. Fix the previous networks and bootstrap the training of RNN decoder on UniM-Poem.
4. Further train the RNN decoder on MultiM-Poem.

4.2 Poem Generation Comparison

To verify the effectiveness of the proposed method, a qualitative comparison is made to show the effects of the model with or without discriminator. As described in Section 3, we first use a image

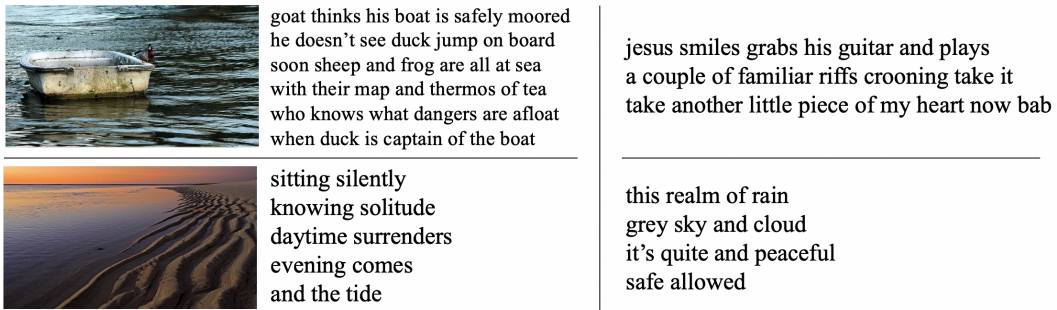


Figure 2: Example of MultiM-Poem dataset(left) and UniM-Poem dataset(right)

feature extraction model combined with LSTM decoder (**PG-LSTM**). A discriminator is then added to our LSTM model as a guide to the poem generation (**PG-LSTM-GAN**).

[14] shows overlap-based evaluation metrics, such as BLEU and METEOR, have little correlation with real human feelings. Thus, we conduct a qualitative comparison to evaluate our method. We compare our results with four baseline models. The models of **Show and Tell** [17], **SeqGAN** [19] and **Regions-Hierarchical** [11] are selected due to their state-of-the-art results in image captioning or paragraphing. And **I2P-GAN** [13] is a state-of-the-art image-to-poem model incorporating deep coupled visual-poetic embedding model and RNN-based adversarial training with multi-discriminators as rewards for policy gradient. The comparison is based on [13] where the baseline models are trained and tested on the same dataset as we use.

One example of result is shown in Table 1, with keywords highlighted by red fonts. As shown in the table, multiple CNNs (**I2P-GAN**, **PG-LSTM**, **PG-LSTM-GAN**) can actually help to generate more image-related poetry as shown by red fonts. The **Regions-Hierarchical** model emphasizes thematic coherence between sentences, while many human poems involve multiple themes or use different symbols for one theme. **SeqGAN** shows that, compared with only the CNN-RNN model, the use of adversarial training for poetry generation has certain advantages, but there is a lack of new concept generation in poetry. We can see that our models perform better on image relevance, especially in sentiment and scene than traditional image captioning and paragraphing models. For example, "sun" can represent "life", "dream", "new day", and "love" in natural association. However, in terms of sentence coherence, our models are weaker than other models. Indeed, we do see the sentence coherence can be improved by applying discriminators when we compare **PG-LSTM** and **PG-LSTM-GAN** in this table as well as in Figure 3, where the model without discriminator outputs some keywords randomly combined together without coherence. More examples generated by **PG-LSTM-GAN** can be seen in Table 2.

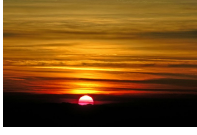
	Regions-Hierarchical [11]	Show and tell [17] (3CNNs)	SeqGAN [19]
	the sun was shining on the sea the waves are crashing in the light's spilled out of heaven and flowing growing the sun is warm and slow	i will find a little bird that shivers and falls the morrow and every day of night and seas immortal night i know that all the world shall be alone and the wild wild horses the women of the great city of the sea and i like to be a jellyfish i will never find a way	the sun is shining on the sea shining on the wind and a sudden green and round the little little boy they said and look at the little noises the with a with a coffee a silver penny a huge a drum steer the spider and you
	I2P-GAN [13]	PG-LSTM	PG-LSTM-GAN
	the sun is singing in the forest wind and let us go to the wind of the sun let the sun be free let us be the storm of heaven and let us be the slow sun we keep our own strength together we live in love and hate	young mountain the a ; no ai a myself inform do life ; ends ; no departure days ; dream worke to to that shod so ; line hes ; not mocked	a swarm of light ; agate and hushed ; hidden from the light ; it has the furthest old ; was nestled in a told ; that new year after day ; moment whether to is love ; what can kill love ; by dreams ;

Table 1: Comparison of our models and baselines based on [13]



PG-LSTM-GAN	PG-LSTM
	glass guilt in the tonight the colour make an starry still ; until lover my stay such ; cold this twist to studs my at which ; bird net want beauty i swoop ; experience veil hands where
	rich and the ; if of dark that even since medicine rich morning day ; i whisper this ; of i frozen life flow and swim
beyond the water ; into the sky jewels ; thin glue ; deep in clouds roads shine off skies	
light of blue little glance ; i think i believe i 'll never share ; from the world for sugar of me ; when i have got junctions leader ; i am a life i see	

Figure 3: Comparison of **PG-LSTM** and **PG-LSTM-GAN**





	<p>oh the song of the night ; above the cool sky ; there is a thousand years ; that winter afternoons ; and in the sun ; we 'll see that drop and fears ; which is the feeling ; let us lead ; all the arbour 's fly ; who do not be fallen ; you 're strong for you ; weather leaf</p>		<p>there is a dream ; there is a smile on the grass ; up in the patch ; leaves upwelling ; past ; then the sun ; starts to the trees ; more than air ; to movement ; with its own ground ; look around ; springs tears ; beautiful and snow</p>
	<p>the mind that a mourning ; somewhere you shall not claim ; in eventide it will be ; your life soars its little controlling ; we can not succeed ; as though .</p>		<p>an echo of its spirit ; of place moving celebration ; the forgotten road suspended ; nature tumultuous voice ; fill the new horns</p>

Table 2: Example of poems generated by PG-LSTM-GAN

5 Conclusion

Generating poems from images is more challenging than generating narrative captions. In this project, we trained a model to generate poems from any image. Our model includes image feature extraction and embedding that are jointly trained with parallel poems, an LSTM decoder to generate poetic language from aligned embedding, and an adversarial discriminator to refine the generated texts. Our model are trained based on a small parallel image-poem dataset, and a large poem-only dataset. Experiment results show that our models successfully generated reasonable poems from given images. The comparison with other state-of-the-arts methods illustrated that our method is powerful and effective.

Contribution

Zhaoyang Li: Implemented the Feature Alignment model. Refined the RNN decoder by introducing adversarial training. Implemented beamsearch for sampling. Integrated different models and conducted training and testing.

Haipeng Li: Contributed on RNN decoder code including data loader, RNN model first version, poem sampling code first version. Contributed on the experiments part of the report.

Ge Shi: Incorporated pre-trained model on scenic features; made the poster; contributed on the approach section and proofreading of the report.

Xi Yang: Contributed to sentimental model of image information extraction, part of poster creation, introduction, related work, conclusion sections of the report.

References

- [1] Tseng-Hung Chen, Yuan-Hong Liao, Ching-Yao Chuang, Wan Ting Hsu, Jianlong Fu, and Min Sun. Show, adapt and tell: Adversarial training of cross-domain image captioner. In *ICCV*, pages 521–530, 2017.
- [2] CrowdFlower. Image Sentiment Polarity - dataset by crowdflower — data.world. <https://data.world/crowdflower/image-sentiment-polarity>, 2016. Online; retrieved Dec 2018.
- [3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [4] Marjan Ghazvininejad, Xing Shi, Yejin Choi, and Kevin Knight. Generating topical poetry. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1191. Association for Computational Linguistics, 2016.
- [5] Jing He, Ming Zhou, and Long Jiang. Generating chinese classical poems with statistical machine translation models. In *AAAI*, 2012.

- [6] 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*, pages 770–778, 2016.
- [7] Jack Hopkins and Douwe Kiela. Automatically generating rhythmic verse with neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 168–178. Association for Computational Linguistics, 2017.
- [8] Hakan Inan, Khashayar Khosravi, and Richard Socher. Tying word vectors and word classifiers: A loss framework for language modeling. *arXiv preprint arXiv:1611.01462*, 2016.
- [9] 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*, pages 3128–3137, 2015.
- [10] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.
- [11] Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 3337–3345. IEEE, 2017.
- [12] Juntao Li, Yan Song, Haisong Zhang, Dongmin Chen, Shuming Shi, Dongyan Zhao, and Rui Yan. Generating classical chinese poems via conditional variational autoencoder and adversarial training. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3890–3900. Association for Computational Linguistics, 2018.
- [13] Bei Liu, Jianlong Fu, Makoto P. Kato, and Masatoshi Yoshikawa. Beyond narrative description: Generating poetry from images by multi-adversarial training. In *Proceedings of the 26th ACM International Conference on Multimedia, MM ’18*, pages 783–791, New York, NY, USA, 2018. ACM.
- [14] Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*, 2016.
- [15] Ofir Press and Lior Wolf. Using the output embedding to improve language models. *arXiv preprint arXiv:1608.05859*, 2016.
- [16] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc., 2014.
- [17] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition*, 2015.
- [18] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France, 07–09 Jul 2015. PMLR.
- [19] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In *AAAI*, pages 2852–2858, 2017.
- [20] Xingxing Zhang and Mirella Lapata. Chinese poetry generation with recurrent neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 670–680. Association for Computational Linguistics, 2014.
- [21] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.