Image to Modern Chinese Poetry Creation via a Constrained Topic-aware Model

LINGXIANG WU and MIN XU, School of Electrical and Data Engineering, University of Technology Sydney, Australia
SHENGSHENG QIAN, National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences and University of Chinese Academy of Sciences, China
JIANWEI CUI, Xiaomi Corporation, China

Artificial creativity has attracted increasing research attention in the field of multimedia and artificial intelligence. Despite the promising work on poetry/painting/music generation, creating modern Chinese poetry from images, which can significantly enrich the functionality of photo-sharing platforms, has rarely been explored. Moreover, existing generation models cannot tackle three challenges in this task: (1) Maintaining semantic consistency between images and poems; (2) preventing topic drift in the generation; (3) avoidance of certain words appearing frequently. These three points are even common challenges in other sequence generation tasks. In this article, we propose a Constrained Topic-aware Model (CTAM) to create modern Chinese poetries from images regarding the challenges above. Without image-poetry paired dataset, we construct a visual semantic vector to embed visual contents via image captions. For the topic-drift problem, we propose a topic-aware poetry generation model. Additionally, we design an Anti-frequency Decoding (AFD) scheme to constrain high-frequency characters in the generation. Experimental results show that our model achieves promising performance and is effective in poetry's readability and semantic consistency.

CCS Concepts: • **Computing methodologies → Natural language generation**;

Additional Key Words and Phrases: Image captioning, poetry generation, semantic consistency

ACM Reference format:
Lingxiang Wu, Min Xu, Shengsheng Qian, and Jianwei Cui. 2020. Image to Modern Chinese Poetry Creation via a Constrained Topic-aware Model. *ACM Trans. Multimedia Comput. Commun. Appl.* 16, 2, Article 53 (May 2020), 21 pages.
https://doi.org/10.1145/3381858

1 INTRODUCTION

Recently, artificial creativity has attracted increasing research attention in the field of multimedia understanding [30, 31, 50] and artificial intelligence [5, 9, 39, 40, 44]. Researchers have been

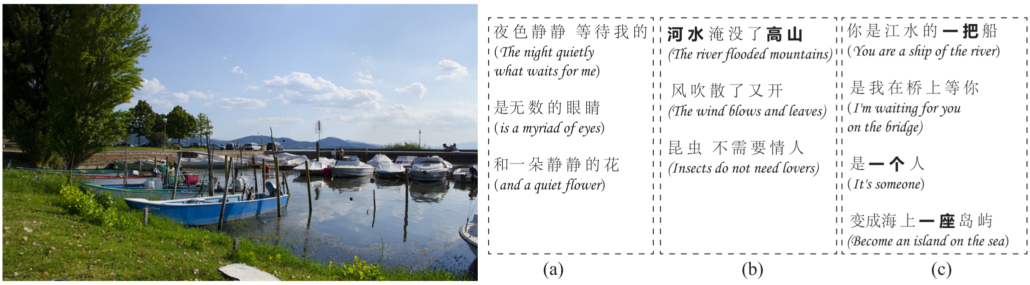
Authors' addresses: L. Wu and M. Xu, School of Electrical and Data Engineering, University of Technology Sydney, 15 Broadway, Ultimo, NSW, 2007, Australia; emails: Lingxiang.Wu@student.uts.edu.au, Min.Xu@uts.edu.au; S. Qian, National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences and University of Chinese Academy of Sciences, 95 Zhongguancun Road, Haidian District, Beijing, China; email: shengsheng.qian@nlpr.ia.ac.cn; J. Cui, Xiaomi Corporation, 68 Qinghezhong Street, Haidian District, Beijing, China; email: cuijianwei@xiaomi.com.

1551-6857/2020/05-ART53 $15.00
https://doi.org/10.1145/3381858

exploring machine capability in human-level creative products [22, 34, 54, 55] like poetry, story, music, painting, and so on. Poetry is regarded as an advanced form in linguistic communication as its expressive form in aesthetics and semantics. In this article, we focus on creating modern Chinese poetry with visual inspirations, which has rarely been investigated. Our work can be applied to various scenarios, e.g., generating personalized poems for photo storage/sharing platforms like Google Photo, Instagram, and so on. It can also enrich the function of chatbot system on phone/tablet/PC terminals. Moreover, automatically creating modern Chinese poetry from images is a challenging task in artificial creativity, which interacts with computer vision as well as natural language processing.

Existing works on automatic poetry generation are mostly rule-based approaches or statistical ones. The rule-based approaches [26, 27, 36, 46] focus on the form of words, characters, rhythms and templates. Netzer et al. [26] proposed to generate Japanese poetry Haiku from a seed word using association norms. The algorithm in Reference [27] created Portuguese poems by filling in sentence templates. However, the rule-based approaches are inappropriate for our task as modern Chinese poetry is mostly written in free verse, which does not follow consistent patterns. The statistical approaches [8, 41, 42, 53, 54] learn to generate poems by extracting statistical patterns in the existing poetry corpus. In Reference [54], classical Chinese poetry can be generated by a Recurrent Neural Network language model. In Reference [41], Wang et al. proposed to create classical Chinese Song iambics with an attention-based sequence-to-sequence model, in which the first sentence should be provided as a cue sentence. Even though these algorithms show promising results in poetry generation, few effort has been attempted on the modern Chinese poetry, which is significantly different from the classical Chinese poetry. Importantly, these algorithms cannot explore the semantic consistency between poems and images. Moreover, existing algorithms fail to deal with a practical problem, i.e., the frequent occurrence of certain words. In fact, some words like “I,” “you,” “one,” and “dream” occupy a large ratio in the modern Chinese poetry corpus. The imbalanced training data make themselves frequently appear in the generation results, which may undermine the readability of generated poems. Therefore, it is critical and essential to explore a method to convey the visual semantics and tackle the high-frequency-word problem for poetry generation.

Recently, some keyword-based algorithms [42, 47, 49] are introduced to handle semantic topics in sequence generation. In Reference [42], four keywords are arranged in an order manually for classical Chinese poetry, and each keyword is assigned with a line as a subtopic. In Reference [47], Xing et al. proposed to generate interesting responses for chatbots by incorporating topic words in the sequence model [35]. Therefore, it is reasonable to convey the visual semantics with topic keywords in the poetry generation. However, existing algorithms still cannot deal with the following challenges in our task: (1) The generation may perform semantic inconsistency with the given image. As shown in Figure 1(a), the generated example is defective due to the irrelevance with the image. (2) The generation may suffer the topic-drift problem. The “topic drift” is defined as the scenario that the generation deviates from the given topic gradually. In other words, the generator “forgets” the visual semantics gradually. One example is shown in Figure 1(b), where only the first line is related to the visual semantics. We refer to that in the conventional sequence model, such as variations of Seq2Seq [35], the encoded vector is only used to provide initial states, thus the generation may deviate from the topic induced in the first line. (3) The generation may suffer the high-frequency-word problem as mentioned above in modern Chinese Poetry, which undermines the readability. The poem shown in Figure 1(c) performs poor readability as the word “One” (shown in bold) repeatedly appears among four lines. These challenges are worth investigating, since they are common and practical in natural language generation.



夜色静静 等待我的
(*The night quietly
what waits for me*)

是无数的眼睛
(*is a myriad of eyes*)

和一朵静静的花
(*and a quiet flower*)

(a)

河水淹没了**高山**
(*The river flooded mountains*)

风吹散了又开
(*The wind blows and leaves*)

昆虫 不需要情人
(*Insects do not need lovers*)

(b)

你是江水的**一把**船
(*You are a ship of the river*)

是我在桥上等你
(*I'm waiting for you
on the bridge*)

是**一个**人
(*It's someone*)

变成海上**一座**岛屿
(*Become an island on the sea*)

(c)

Fig. 1. Three challenges in image-poetry generation: (a) Semantic inconsistency between the image and the poetry; (b) topic drift in the subsequent lines (relative words are shown in bold); (c) frequent occurrence of certain words (shown in bold).

In this article, we identify three major challenges in image-to-modern-Chinese-poetry creation. To tackle these three challenges, we propose a Constrained Topic-aware Model (CTAM). To ensure the semantic consistency between images and poems, we utilize an image caption model as a bridge. A visual semantic vector is constructed with key components in the caption, and it serves as the poetry topic. To deal with the topic-drift problem, we propose a topic-aware poetry generation model, which contains two LSTMs at each step. We use a temporal LSTM to transmit sequential Chinese characters, and use a depth LSTM to recall the semantic vector selectively. To constrain the occurrence of high-frequency characters, we propose an Anti-frequency Decoding (AFD) scheme based on Mutual Information (MI). The pipeline of our approach is briefly illustrated in Figure 2. There does not exist any image-poetry (modern Chinese poetry) dataset for end-to-end training. Matching images with poems is diversely subjective. Thus, we exploit image captioning as a bridge to convey visual semantics. The image description is translated into a Chinese sentence, from which a semantic vector is formed as the poetry topic. The captioning model is typically optimized with cross-entropy loss, which cannot correlate well with human assessments. To generate captions with rich semantics like human creating, we apply the captioning model augmented by Reinforcement Learning [32, 33, 43] in this task. Then, through the topic-aware poetry generation model and anti-frequency decoding, an appropriate poem can be created. In the experiment, the qualitative evaluation and quantitative evaluation clearly demonstrate that our approach can effectively perform semantic consistency and proper readability.

The main contributions of our work can be summarized as:

- This is a pioneer study on modern Chinese poetry creation from visual information. We figure out three major challenges in this task: semantic inconsistency, topic drift, and word re-appearance.
- We propose a Constrained Topic-aware Model for image-to-poetry creation, where we introduce visual semantic vectors to ensure semantic consistency between images and poems, we propose a topic-aware poetry generation model to recall visual topics selectively and prevent the topic-drift problem, and we propose the Anti-frequency Decoding scheme based on MI to constrain the occurrence of particular characters.
- Qualitative and quantitative experiments demonstrate the effectiveness of our approach, and our design leads to dramatic improvement on readability and semantic consistency.

The rest of the article is structured as follows. In Section 2, some previous works on image captioning and poetry generation are reviewed. In Section 3, we describe our approach in details. The experiment is discussed in Section 4.

2 RELATED WORK

2.1 Image Captioning

Image captioning is to automatically generate a sentence description given an image, which attracts research efforts from both computer vision and natural language processing communities. Approaches proposed for image captioning can be loosely summarized into three categories: retrieval-based approaches, template-based approaches, and neural-network-based approaches.

Retrieval-based methods [6, 12, 28] retrieve the most relative sentence by a designed similarity metric. In Reference [12], Hodosh et al. collected an image-caption dataset consisting of 8,000 images, where each image is paired with five sentences. Kernel Canonical Correlation Analysis was employed to associate images and captions. The limitation is that these retrieval-based methods cannot describe images with the new components as they only feed back existing captions. The performances are usually affected by the sizes of datasets. Template-based approaches [16, 19] generally extract phrases and compose captions with linguistic constraints. For example, in Reference [19], n-gram phrases in designed patterns were retrieved from web data based on the components in images. The optimal compatible set of phrases was then composed using dynamic programming. As these templates are manually designed, this kind of methods usually lacks syntactic variability.

Neural-network-based methods [38, 45, 48, 51], especially the encoder-decoder framework, had attracted great research attention recently as the ability to train visual data and text data in an end-to-end manner. In Reference [38], high-level visual features were encoded by a Convolutional Neural Network. A Recurrent Neural Network composed of LSTM received the image information through the initial interactions and predicted subsequent words given the previous word. It is straightforward, but it may lose rich information in images. To deal with it, in Reference [48], attention mechanism was exploited in the image captioning. Yang et al. [51] extended the attentive encoder-decoder framework with a review network between the encoder and the decoder. These algorithms introduced extra networks for attention mechanism. Besides convolutional neural network features, semantic features such as image attributes predicted by a separate system were also used in Reference [52].

Different from the above-mentioned methods, in this work, we use image captioning merely as a bridge towards poetry generation, since captions contain more abundant semantic descriptions such as nouns, adjectives and prepositions compared with object recognition. We exploit a neural network model to generate captions for images.

2.2 Poetry Generation

Poetry generation is an interesting task in intelligent computational creativity. A variety of approaches have been proposed on this topic, while most of them are inspired by text information rather than visual semantics. Existing work on poetry generation can be roughly divided as rule-based approaches and statistical approaches.

The rule-based approaches generally focus on the form of words or characters, and they generate poems based on patterns or templates. [26, 36, 46] were proposed to generate Japanese poetry, Haiku/Renku, by searching phrases that match the syntactic patterns and theme words. The algorithm in Reference [36] started with a user-supplied seed word and generated poems based on syntactic constraints. The algorithm in Reference [26] generated Haiku from a seed word using association norms. However, all of them depend on the sensitivity of the seed phrase. Oliveira [27] proposed to generate Portuguese poetry by filling in sentence templates. The algorithm in Reference [4] also used templates to construct poems according to constraints on rhyme, meter, stress, sentiment, word frequency and word similarity. These approaches can benefit user-interactive generation. However, the rule-based approach is not suitable for modern Chinese poetry generation, as free verse is the dominant form. A fixed template is not available for modern Chinese poetry.

More recently, statistical approaches received more research attention by constructing a language model to extract statistical patterns in an existing poetry corpus. In Reference [54], given keywords, Chinese quatrains can be generated by a Recurrent Neural Network language model. The first line was generated based on keywords, while the subsequent lines were generated based on all previously generated lines. In Reference [41], Wang et al. applied an attention-based sequence-to-sequence model to generate Chinese Song iambics, in which the first sentence should be provided as a cue sentence. These two approaches share the limitation that the subsequent lines may deviate from the semantic topic inducted at the first line. To address it, Zhang et al. [53] proposed a memory-augmented neural network to consider innovation in Chinese classical poems. In Reference [42], to generate quatrains, a keyword is assigned to each line. The problem is that the keyword orders have to be set manually, which limits the poem flexibility and results in incoherency across lines. In Reference [13], a phonetic-level neural network model and a constrained character-level neural network model were proposed for rhythmic poetry in a variety of forms in English.

Recently, some works generated styled descriptions of images, which is related to image-poetry generation somehow. In StyleNet [7], Gan et al. proposed to generate humorous or romantic descriptions for images using a standard image-caption dataset and an external monolingual text dataset. In SentiCap [23], Mathews et al. proposed to generate captions with positive or negative sentiments. This is achieved by two labeled datasets: a standard image-caption dataset and a dataset containing captions with sentiments. This supervised transforming method is interesting, but it is expensive and difficult to scale up.

Different from the above-mentioned methods, in this work, we generate modern Chinese poems from images. This work directly connects visual semantics with modern Chinese poetry generation. The existing work [49] generated Chinese quatrains from images using a memory-augmented network. However, an image-poetry paired dataset must be provided. Additionally, it ignored the semantic gaps between image objects and classical Chinese poetry contents. In Reference [21], Liu et al. proposed to generate English free verse from images in an end-to-end fashion. While, a large paired dataset should be collected and annotated by human annotators. In our work, there is no image-poetry (modern Chinese poetry) dataset. We utilize image caption as a bridge and keep the semantic between images and poetries consistently. Moreover, due to the variant form of modern Chinese poetry, constructing a language model is more challenging in our work.

3 CREATING POETRIES FROM IMAGES

3.1 Overall Framework

Given an image I, our goal is to create a modern Chinese poem W, which can be represented as a sequence of Chinese characters $W = [w_0, w_1 \dots w_N]$. To achieve this, we construct a generation model for $p(w_t|I, w_0, \dots w_{t-1})$. The framework of our approach is depicted in Figure 2, which contains three major steps: visual semantic vector construction, topic-aware poetry generation, and anti-frequency decoding. The advantages of each processing step are listed as follows:

- **Visual Semantic Vector Construction:** We choose the image caption as a bridge to convey the poem topics. First, compared with a set of object categories, the caption contains much more rich descriptive information. Second, compared with References [45] and [20], which generate Chinese quatrains or English free verse from images on paired dataset, we do not need an image-poem dataset.
- **Topic-aware Poetry Generation:** Compared with the conventional sequence model, which is constructed with LSTM and the visual context is only injected through the initial state, our topic-aware generator uses the temporal LSTM (tLSTM) to transmit the sequence

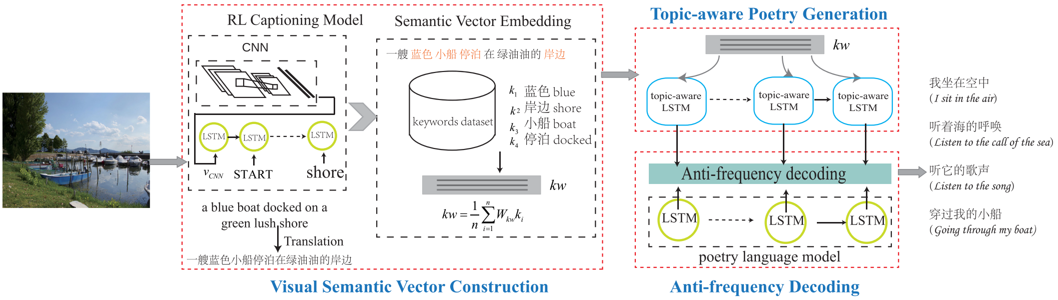


Fig. 2. The framework of our proposed CTAM.

of Chinese characters, and it uses the depth LSTM (dLSTM) to recall the semantic vector at each step. In this way, the proposed generator provides a unique way to involve keyword topics adaptively and continually.

- **Anti-frequency Decoding:** Compared with the general decoding scheme such as greedy sampling or beam search, the proposed scheme can penalize frequent words during the decoding stage. This is useful for some imbalanced dataset in which some words/characters frequently appear.

Specifically, we use the image caption $C = [c_0, c_1 \ldots c_M]$ as a bridge between images and poems, which encodes visual semantics and provides the topics for the generation. To present the visual semantics, a semantic vector kw is constructed by keywords $k_1, k_2 \ldots k_n$ in an English-Chinese translated caption. Then, the semantic vector kw is injected into our proposed topic-aware poetry generation model, which can deal with the topic-drift problem in the Recurrent Neural Network. With topic-aware LSTMs, we get $p(w_t|kw, w_0, \ldots w_{t-1})$. Last, in decoding, an anti-frequency decoding score, s_t, is proposed to inhibit the frequently appearing characters such as "I," "Love." We list key notations in Table 1.

3.2 Visual Semantic Vector Construction

To create poetries related to the given image, we need to mine semantic information from the visual image first. As there is no available image-poetry paired dataset, we exploit the image caption as a bridge in this article, and then we construct a semantic vector through keywords in the caption.

3.2.1 Captioning Model. Image captioning is an appropriate bridge connecting computer vision and natural language processing. Different from traditional computer tasks such as object recognition that only recognizes major objects in an image, image captioning can generate a natural language sentence to describe the full image. A caption contains ample semantic information including objects, adjectives, verbs even prepositions.

We use a CNN-RNN model [38] to generate captions. The given Image I is encoded with a Convolutional Neural Network (CNN) as v_{CNN}. The encoded features are then decoded into a sequence of words by a Recurrent Neural Network (RNN). The objective function is typically to minimize the cross-entropy loss,

$$L(\theta) = -\sum_{t=1}^{M} \log p\left(c_t | v_{CNN}, c_0 \ldots c_{t-1}; \theta\right), \tag{1}$$

where θ are the parameters to be learnt.

However, the typical objective function cannot correlate well with human assessments. The non-differentiable metric CIDEr [37] can measure the captioning quality using human consensus.

Table 1. Listing of Notations

Notation	Description
I	the input image
$W = [w_0, w_1 \dots w_N]$	the generated poem
$C = [c_0, c_1 \dots c_M]$	the generated image caption
k_i	a keyword in the translated caption
kw	the semantic vector
$G = (V, E)$	text graph in the TextRank algorithm
$WS(V_i)$	the TextRank score for vertex i
w_{ji}	weight of an edge in the text graph
d	a damping factor in the TextRank algorithm
h_t	a hidden state in RNN
i_t, f_t, o_t	input/forget/output gates in LSTM
C_t	a memory cell in LSTM
v_t	the frequency of the character w_t in corpus
s_t	the decoding score
α	the frequency threshold in the decoding scheme
λ	the penalty hyperparameter in decoding scheme
W_{kw}	the keyword embedding matrix to be learnt
$U_{i/f/o/C}, W_{i/f/o/C}$	the parameters in LSTM to be learnt

Recently, a Reinforcement Learning (RL) algorithm, REINFORCE [43], was introduced in the recurrent neural network to deal with the gradient of the non-differentiable function [32]. It can optimize the gradient of the expected reward (the CIDEr score in this article) by sampling from the model during training. To get captions with rich semantic information like human creating, we exploit an RL-captioning model in this task. A self-critic algorithm [33] is exploited as it avoids estimating the reward signal through the self-critics. The alternative objective function is to minimize the negative expected reward, and the expected gradient of the reward function can be computed as follows:

$$L(\theta) = -\mathbb{E}_{c^s \sim p_\theta} r\left({c_0}^s, {c^s}_1 \dots {c_M}^s\right), \tag{2}$$

$$\Delta_\theta L(\theta) \approx -\left(r(C^s) - r(C^g)\right) \Delta_\theta \log p_\theta(C^s), \tag{3}$$

where $r(C^s)$ is the reward of the caption sampled from the model, and $r(C^g)$ is the reward of the caption obtained by greedy sampling.

Since there is no available large-scale Chinese image-caption paired dataset, we use MSCOCO [3] dataset to train the captioning model, and then we translate the English descriptions into Chinese with Baidu Translation API.[1] Different from classical Chinese poetry, modern Chinese poetry is mostly written with modern vernacular, which can be easier to correspond to the translated captions. Generally, the captioning model trained on MSCOCO can describe most of the images.

3.2.2 Semantic Vector Embedding. We utilize keywords in the translated caption to form a visual semantic vector. First, we collect a poetry keyword dataset. Then, the Chinese caption is segmented into words or phrases. Relative words/phrases in the caption are selected as visual keywords by retrieving in the poetry keywords dataset. At last, the visual keywords are represented as word embedding vectors and form a semantic vector.

[1] https://fanyi.baidu.com/.

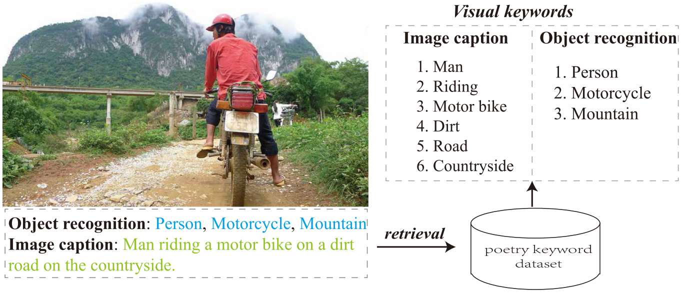


Fig. 3. Comparison on keywords obtained from two different visual models.

To construct the poetry keyword collection firstly, we extract keywords from each poem by an unsupervised method TextRank [24]. The TextRank algorithm is a graph-based ranking algorithm to evaluate the importance of words. Text is represented as a graph $G = (V, E)$, and words can be added as vertices in the graph. Edges are added between two words based on the co-occurrence. The TextRank score is iterated as

$$WS(V_i) = (1 - d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j), \tag{4}$$

where w_{ji} is the weight of the edge, and d is a damping factor that is set to 0.85 in our experiment. The algorithm is iterated till converge, and the vertices are sorted based on their scores. Words with higher scores are selected as keywords in the corresponding text.

In Figure 3, we show keywords obtained from two different visual models. It is clear that the caption provides rich information, while object recognition only provides three object names. It is reasonable to exploit the captioning model for visual information interpretation.

Admittedly, some styled image descriptions [7, 23] are more attractive and contain more emotional words. However, in our approach, we use image-caption data to achieve visual consistency, and then we train a generation model on poem corpus to achieve poetic imagination. Intuitively, we would like to avoid the biases towards a large number of emotional topic words in our task. For instance, given an image with a *moon*, we prefer the poem related to the *moon* rather than only related to *loneliness*. We argue that the basic adjectives and verbs generated by our captioning model contain sentiments somehow.

To find the proper visual keywords, we segment the caption into words, then retrieve each word in the poetry keyword dataset. To encode the visual semantic information, we construct a semantic vector by the selected visual keywords. A pre-trained word2vec [25] model is utilized then. Each keyword is represented as a word embedding vector, and the semantic representation is formed as an average of the keyword embedding vectors:

$$kw = \frac{1}{n} \sum_{i=1}^{n} W_{kw} k_i. \tag{5}$$

W_{kw} is the pretrained word2vec embedding matrix. In this way, the visual information is encoded as a semantic vector, and will be decoded into a piece of poetry relative to the image.

3.3 Topic-aware Poetry Generation

To generate a poem from the semantic vector, a character-level RNN generation model is exploited then. We regard the semantic vector as the topic for a poem and input it into the RNN generation model. The output for the RNN is a poem that consists of a sequence of characters. The generation model is trained by cross entropy loss:

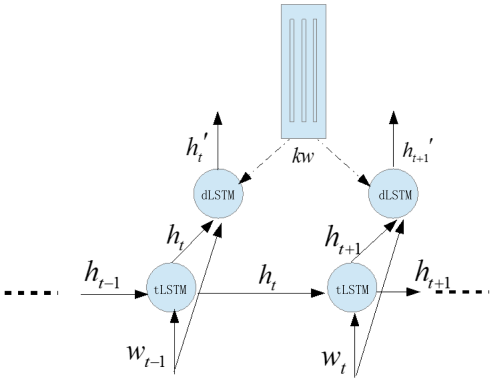


Fig. 4. The topic-aware LSTM details.

$$L(\delta) = -\log \mathrm{P}_\delta (W = w|K = kw), \tag{6}$$

where

$$\mathrm{P}_\delta (W = w|K = kw) = \prod_{t=1}^{N} p(w_t|w_1, \ldots w_{t-1}, kw; \delta). \tag{7}$$

However, in the conventional sequence model [35], the encoded vector would be injected by the initial states in the decoder. This may induce the RNN "forget" the initial input along the sequence generation. The generated text may deviate from the encoded vector, which induces a topic-drift problem. To tackle this problem and maintain the semantic consistency, we propose a novel topic-aware model for poetry generation.

The topic-aware model is inspired by GridLSTM [14], which contains two dimension LSTMs: a temporal LSTM and a depth LSTM. In our method, we not only inject the semantic vector through the initial state but also inject it through the dept LSTM at every step. We use these two LSTMs to make the model "remember" topics consistently and adaptively. Specifically, We use the temporal LSTM (tLSTM) to transmit the sequence of Chinese characters, and we use the depth LSTM (dLSTM) to recall the semantic vector at each step. At each time step, two LSTM computations are involved, as presented in Figure 4:

$$h_t = tLSTM(w_{t-1}, h_{t-1}, C_{t-1}), \tag{8}$$

$$h_t' = dLSTM(w_{t-1}, h_t, kw), \tag{9}$$

$$p(w_t|kw, w_0, \ldots w_{t-1}) = f(h'_t), \tag{10}$$

where w_{t-1} is the representation of the character at $t-1$. f is a Softmax function with parameters W_s and b_s, softmax$(W_s h'_t + b_s)$.

The tLSTM works as the conventional LSTM [11] to transmit the sequence of Chinese characters:

$$\begin{aligned}
i_t &= \sigma(U_i h_{t-1} + W_i w_{t-1} + b_i), \\
f_t &= \sigma(U_f h_{t-1} + W_f w_{t-1} + b_f), \\
o_t &= \sigma(U_o h_{t-1} + W_o w_{t-1} + b_o), \\
\tilde{C}_t &= \tanh(U_C h_{t-1} + W_C w_{t-1} + b_C), \\
C_t &= f_t \odot C_{t-1} + i_t \odot \tilde{C}_t, \\
h_t &= o_t \odot \tanh(C_t),
\end{aligned} \tag{11}$$

where i_t, f_t, o_t are the input gate, forget gate, output gate, respectively. C_t is the memory cell. $W_{i/f/o/C}$, $U_{i/f/o/C}$ and $b_{i/f/o/C}$ are parameter matrices to be learned. σ is the sigmoid activation function where $\sigma(x) = 1/(1+\exp(-x))$. $\odot$ denotes the product with a gate value.

The dLSTM can admit the semantic vector, kw, through the previous *memory*. Computation details in the dLSTM are as follows:

$$\begin{aligned}
{i_t}' &= \sigma\left(U_i h_t + W_i w_{t-1} + b_i\right),\\
f_t' &= \sigma(U_f h_t + W_f w_{t-1} + b_f),\\
\tilde{C}_t' &= \tanh\left(U_C h_t + W_C w_{t-1} + b_C\right),\\
{C_t}' &= {f_t}' \odot kw + {i_t}' \odot \tilde{C}_t',\\
{o_t}' &= \sigma\left(U_o h_t + W_o w_{t-1} + b_o\right),\\
{h_t}' &= {o_t}' \odot \tanh({C_t}').
\end{aligned} \tag{12}$$

In the dLSTM, the semantic vector, kw, is constantly recalled as the previous memory. In the depth dimension, the previous hidden state is h_t, and the input vector is w_{t-1} here.

The LSTM itself is a useful net structure. According to Reference [11], the net can use input gate to decide when to keep or override information in the memory cell, and the output gate to decide when to access the memory cell and when to prevent other units from being perturbed by the memory cell. In Reference [14], the LSTM mechanism is described as a "memory and implicit attention system, whereby the signal from some input can be written to the memory vector and attended to in parts across multiple steps by being retrieved one part at a time." In our situation, we propose to involve the topic semantic vector through the depth LSTM. Our method can involve the keyword topic continually and adaptively. First, compared with the conventional sequence model, which injects the conditional information only as the RNN initial state, we utilize the depth LSTM to involve the semantic vector at each step along with the generation. Second, through the LSTM mechanism, the semantic information can be involved adaptively. Intuitively, the input gate controls the extent to which a new value flows into the cell, and the forget gate controls the extent to which a value remains in the cell. As shown in Equation (12 (4)), the semantic vector kw is recalled as the previous memory and it interacts with the forget gate value. At each step, the memory cell ${C_t}'$ involves new information to update and drop certain semantic information that is selected to be forgotten. Thus, at each step, the semantic information is integrated in the network considering the current content such as h_t and w_{t-1}, which handles the poetry topic adaptively.

In our method, the caption model is exploited to achieve visual consistency, while the generation model can achieve poem imagination. As the generation model is trained on the poem corpus, it faces poetry language constraints. The poetry generation model can always generate poetic style sentences that contain emotional words. Moreover, with the topic-aware mechanism, the generator can always generate emotional words related to the given topics. In this way, the generation model can generate emotional words compatibly even though only factual words are used as topics.

We note TA-Seq2Seq [47] and GridLSTM [14] have some connections with our work. The differences are as follows. In the topic aware sequence-to-sequence (TA-Seq2Seq) [47], the author proposed a joint attention [1] model to leverage topic information in dialogue generation. Differently, we focus on generating poems with visual semantic vectors. Moreover, we use a newly designed depth LSTM mechanism instead of the attention mechanism [1] to "remember" the topics. GridLSTM [14] is a network where the LSTM cells can be arranged in a multidimensional grid. Although our topic aware generation model is inspired by GridLSTM [14], it is different in a large extent. In our work, we utilize the depth LSTM and design a novel data flow for our own task.

Being injected through the previous memory cell, the topic vector gets involved adaptively as the memory cell will be updated considering the topic vector and the newly generated information.

3.4 Anti-frequency Decoding

Generator trained with Equation (6) will assign higher probabilities to Chinese characters frequently appearing in the training corpus. In modern Chinese poetry collection, some words/characters, such like “I,” “heart,” “dream” appear in a higher frequency compared with others. We regard it as an imbalanced data distribution problem, where a dataset exhibits an unequal distribution between its classes. This problem is worth careful investigation as it often appears in real-world scenarios.

To solve this problem, we propose a novel decoding scheme based on Mutual Information (MI) [2]. The Mutual Information of two random variables is a measure of the mutual dependence between the two variables. Maximum Mutual Information (MMI) was introduced as objective functions in speech recognition [2] and dialogue generation [18]. In Reference [18], an MMI-based objective function was explored to generate more diverse, interesting, and appropriate responses in dialogue generation. Thus, it is reasonable to design a decoding scheme with MI.

Our decoding scheme aims to constrain those high-frequency characters and consequently maintain the poetry readability. The mutual information between the sample semantic vector kw and the sample character sequence w in the poetry is

$$\begin{aligned} I\left(W,K\right) &= \log \frac{P\left(K=kw,W=w\right)}{P\left(K=kw\right)P\left(W=w\right)}, \\ &= \log \frac{P\left(W=w|K=kw\right)}{P\left(W=w\right)}, \\ &= \log P\left(W=w|K=kw\right) - \log P\left(W=w\right). \end{aligned} \tag{13}$$

The second term makes a difference. It penalizes characters in the language model $P(W=w)$.

To merely inhibit the high-frequency characters, we take the frequency of Chinese characters in the training corpus into account. We define an alternative decoding score as

$$s_t = \log p\left(w_t|w_1\ldots w_{t-1},kw\right) - \lambda g\left(\alpha,v_t\right)\log p\left(w_t|w_1\ldots w_{t-1}\right), \tag{14}$$

$$g\left(\alpha,v_t\right) = \begin{cases} 1 & v_t \geqslant \alpha \\ 0 & v_t < \alpha \end{cases}, \tag{15}$$

where λ is a penalty hyperparameter, v_t is the frequency of character w_t in the training corpus, and α is the frequency threshold. Thus, high-frequency words in language model $P(W=w)$ will get penalties during decoding.

In our decoding scheme, only the high-frequency characters that appear more than frequency α in the corpus would get penalties. Our model can constrain the high-frequency characters and maintain the readability.

4 EXPERIMENTS

We start with the collected poem dataset and then the implement details. To show the performance of our model, we present the results and analysis with both qualitative and quantitative evaluation.

4.1 Dataset

To train an image-captioning model, we use the MSCOCO [3] dataset, which is provided for Microsoft COCO caption challenge. There are 82,783 images in the published training set, 40,504 in the validation set and 40,775 images in the test set without annotations. Each image in the training set and validation set is annotated with five descriptive English sentences. As in

References [38, 51], we merge all published annotated data, and allocate 5,000 for validation and test split, respectively. The rest 113,287 images are used for training. In this article, the visualized image examples are from MSCOCO.

For the poetry generation, we collected 16,015 modern Chinese poetries on the Internet from the beginning of the 20th Century until now. Most of the poetries are collected from distinguished poets' anthologies, and some are crawled from poetry forums where poetry fans posted their compositions. To enlarge the data collection, we also added 1,189 poetries translated from other languages to Chinese by experts. The dictionary for poetry generation contains 6,194 Chinese characters. The length of poetries vary significantly. Thus, we cut long poetries into a few parts. For each poetry, we extract corresponding keywords as described in Section 3.2.2.

The keywords include nouns, adjectives as well as verbs. "world" (occurrences 2,050), "sky" (occurrences 2,294), "life" (occurrences 2,211), and "sunshine" (occurrences 1,880) are the most common topics in the poetry keyword collection.

4.2 Implementation Details

While training the captioning model, both RNN node size and word embedding size are set as 512. Dropout is 0.5 experimentally. Resnet-101[10] pretrained on ImageNet is used to encode each image. We do not rescale or crop the image. We encode the full image with the final convolutional layer, and then apply average pooling, which results in a 2048-d vector. We stop training the captioning model with cross-entropy loss after 30 epochs, and then we do RL-based training till 70 epochs. To construct the semantic vectors, the keyword embedding size is set as 512 experimentally. Maximal four keywords are extracted from a poem sample. In the poetry generation model, both RNN size and character embedding size are set as 512. The generation model is obtained after 100 epochs of training. To get a variety of poetries instead of a unified one, we sample the characters given the decoding scores. During the Anti-frequency decoding, we only give penalty to the top-100 frequent characters. λ is set as 0.8 in most of our experiment except Section 4.7 Major details are given as below:

- LSTM size, word/character embedding size, keyword embedding size are all set as 512 experimentally.
- LSTM parameters and word/character embedding parameters are initiated by a uniform distribution in $[-0.08, 0.08]$
- Batch size is set as 16, and the dropout rate is 0.5.
- Adam [15] is utilized for stochastic optimization. Learning rate is initially set as 1×10^{-4}.

Our implementation on a single Nvidia GTX 1080 GPU process at a speed of 0.32 second per iteration.

4.3 Evaluation Metrics

Evaluation of poems is a difficult task. As modern Chinese poetry is in free verse and the contents are dramatically diverse, it is hard to approximate the ground truth given the topic words. For instance, with "Love" and "Dream" the generator can create variants of poems. So NLP metrics such as BLEU [29] is not suitable for evaluation here.

Following References [13] and [54], we use human evaluation with three criteria, namely, Semantic consistency, Readability, and Poeticness/Aesthetics:

- Semantic consistency (S): whether the poetry is related/corresponding to the image in semantics.
- Readability (R): whether the poetry is fluent in a single sentence and/or coherent between lines.

- Poeticness/Aesthetics (P/A): Whether the poetry expressed poetically with respect to aesthetics and emotion.

We consider the overall quality of the image-poetry pair through the average of these three scores. We also report the percents of participants who consider the poetries written by human beings.

To conduct a human evaluation, we invited 46 participants to fill in questionnaires[2] through a WeChat platform. There are ten image-poem pairs on each questionnaire. In these ten image-poem pairs, we randomly selected one or two pairs from each model. Each questionnaire contains examples from all type of models. Participants were required to rate each pair on a 1–5 scale (the higher the better) with respect to three criteria, and distinguish whether the poem is written by human beings. Most of the participants are undergraduates or postgraduates, majored in computer science or Chinese language and Literature, from University of Technology Sydney, Beijing Institute of Technology or Peking University. Besides students, we also invited a few professors in computer science community and people engaged in art community. Participants are with an age range from 19 to 52. All of the participants took part in this challenge as volunteers. For the quality of evaluation, we only listed ten pairs on each questionnaire. Not all the 46 questionnaires are same. We designed different sets of questionnaires and asked participants to randomly choose one to fill. We required readers to read the poems patiently and analyse carefully. It would take some time to finish the questionnaire rather than scoring at the first sight roughly.

4.4 Comparative Methods

We compare the proposed **CTAM** with three kinds of methods. (1) We extend two existing generation methods (Sequence model [35] and PPG [42]) into our task. (2) Poems written by Human and chatbot system XiaoIce are included as well. (3) We also evaluate CTAM without AFD. Details of the comparative methods as listed as follows:

Sequence model: The sequence model is built on the sequence-to-sequence framework [35], which contains an encoder and an decoder constructed by conventional LSTMs. According to the captioning model [38], we modified the network to fit our task, where the semantic vector has been encoded and is injected into the first step of the RNN decoder.

PPG: A planning-based poetry generation approach [42] proposed for classical Chinese poems. Four keywords are sorted in an order manually, and a sub-topic keyword is assigned to each line. We discard the rhythmic restriction in the implementation.

XiaoIce: XiaoIce is a Microsoft chatbot system that can generate poems given an image. We randomly select a group of images including images from MSCOCO dataset and some others. For each image, we generate poems with all the models. For XiaoIce, we upload these images to their platform and collect the poems generated.

Human: Human-written poems. Obscure poems written by poets are randomly selected from our self-collected dataset. A relative image is manually matched with the poetry according to the content.

CTAM w/o AFD: A variation of our proposed CTAM, which performs topic-aware poetry generation without anti-frequency decoding.

4.5 Quantitative Evaluation

4.5.1 Human Evaluation. The evaluation results are reported in Table 2. Obviously, poems written by human beings obtain the highest scores across all evaluation criteria. Among the

[2] One example can be accessed through https://www.wjx.top/jq/20790601.aspx.

Table 2. Human Evaluation Results

Methods	S ↑	R↑	P/A↑	Average Score ↑	Written by Human (%) ↑
Sequence model [35]	2.45	2.45	2.95	2.62	39%
PPG [42]	2.64	2.33	2.52	2.50	35%
XiaoIce	2.69	3.12	**3.39**	3.07	**56%**
Human	**3.46**	**3.42**	**3.40**	**3.43**	**64%**
CTAM w/o AFD	2.81	2.90	3.01	2.91	35%
CTAM (ours)	**2.93**	**3.38**	3.24	**3.18**	49%

Table 3. Ablation Study with BLEU-1 (%)

Model	CTAM	CTAM-a	CTAM-d	CTAM-k
BLEU-1↑	9.89	12.20	5.87	9.52

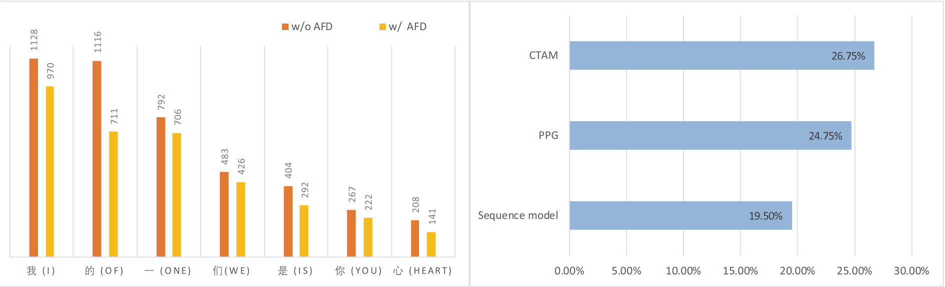
automatically generated poems, poems created by our proposed CTAM obtain the highest score, 3.18, on average. It also can be seen that our approach has an excellent performance in semantic consistency and readability. Specifically, the CTAM w/o AFD and CTAM both outperform the sequence model on semantic consistency. It can be seen that the topic-aware design is able to prevent generation deviating from the semantic vector effectively. On this aspect, our model significantly outperforms the XiaoIce, which is weak on semantic consistency. On readability, our proposed CTAM outperforms the CTAM w/o AFD. We infer that the anti-frequency decoding effectively inhibits the high-frequency words and thus the poems read more fluently. On the Poeticness/Aesthetics, XiaoIce comes second. We noted that most of the poems generated by our method are short (around four lines), while poems generated by XiaoIce mostly contain eight lines (two chunks of four lines and a blank line in the middle). By interviewing some participants, the major reason they gave XiaoIce high scores on the Poeticness/Aesthetics is that poems generated by XiaoIce have a longer form. This might also be the reason why more participants regard poems from XiaoIce as human beings written. 49% participants regard poems generated by our approach as human written. Obtaining 3.18 on average score, our CTAM outperforms all other automatic generators.

4.5.2 Ablation Study with BLEU-1. Since the proposed CTAM contains several key components, we compare variants of CTAM to demonstrate the effect of CTAM: (1) the effect of the topic-aware generator; (2) the effect of the Anti-frequency Decoding strategy; (3) we also investigate the impact of the TextRank keyword extractor. The following CTAM variants are designed:

- CTAM-d: A variant of CTAM with the dLSTM being removed. This model is the same as the LSTM-based sequence model.
- CTAM-a: A variant of CTAM with the Anti-frequency Decoding being removed.
- CTAM-k: A variant of CTAM constructed on another keyword-poem dataset in which the keywords are extracted via the word frequency.

We use BLEU-1 [29] to evaluate the generation quality of our proposed model and the model variants. BLEU-1 analyzes the co-occurrences of 1-grams between the candidate and reference poem. In the 16,015 poems, we use 10% for testing, 10% for validation and 80% for training. The results are shown in Table 3.

(1) *The effect of the topic-aware generator:* To verify the effect of the GridLSTM-based generator, we compare the CTAM and CTAM-d. From Table 3, we can easily observe that the CTAM significantly outperforms the CTAM-d on BLEU-1.



(a) (b)

Fig. 5. Comparison results. (a) The occurrences of seven frequently appearing characters generated by the model w/ and w/o anti-frequency decoding. (b) The percentage of generated poetries containing keywords.

(2) *The impact of the Anti-frequency Decoding:* To see the impact of the proposed decoding strategy, we compare CTAM and CTAM-a. We note that the CTAM-a outperforms CTAM on BLEU-1, but the generations from CTAM-a lack diversity and contain a number of repeated high-frequency words. For instance, there may be three fragments like "a man" in one poem. The CTAM-a gets higher BLUE-1, because these high-frequency words generally appear in the reference text. The human evaluation in Table 2 indicates that the CTAM can generate better poems than the CTAM-a.

(3) *The impact of the keyword extractor:* To see the impact of the TextRank keyword extractor, we compare the CTAM and CTAM-k. The generation quality of CTAM is slightly better than the CTAM-k. Even though the generation quality is comparable, we need consider the keyword interaction with the caption dataset. Among these keywords, there are only 2,164 words in the caption dataset, which is less than the 2,267 intersections with the TextRank-based dataset. Thus, the TextRank algorithm is more suitable for our task.

4.5.3 Comparison between Model w/ and w/o AFD. In this experiment, we compare the performance between model w/ and w/o anti-frequency decoding. We use keywords in the poetry keyword set to construct 400 keyword combinations, each of which contains maximal four keywords. A poem is generated with respect to each combination. Then, we calculate and compare the occurrences of the most commonly appearing characters.

In Figure 5(a), we present the seven most commonly appearing Chinese characters and their occurrences in the generation results.We find that the occurrence of these characters from model w/AFD is clearly lower than that from model w/o AFD. The experimental results indicate that the designed decoding scheme is effective to inhibit characters with high frequency.

4.5.4 Percentage of Generated Poems Containing Keywords. To see how the topic-aware model makes differences from others, we present the percentages of generated poems containing keywords. It is assumed that the more keywords appearing in generated poems, the more semantic consistent. To be specific, we randomly selected 400 keyword combinations to generate poems. Then, we compare the number of poems that contain the input keywords directly. Although some generation may contain words closely relative to the keywords, the comparison can somehow reveal the direct connection between keywords and poems.

As shown in Figure 5(b), 26.75% poems generated by the CTAM contain keywords, while only 19.50% poems generated by the conventional sequence model contain keywords. It demonstrates that the proposed topic-aware generator is more effective than the sequence model to maintain semantic consistency. 24.75% poems from PPG contain keywords directly, which is slightly lower than the CTAM. We note that poems from PPG contain more direct keywords than those from

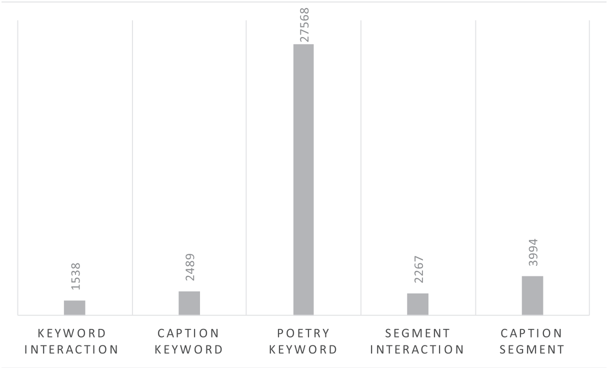


Fig. 6. The number of poetry keywords and caption words.

Table 4. Image-captioning Performance

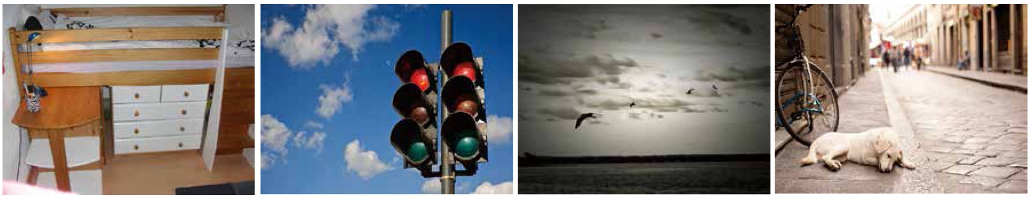
Model	BLEU-4	METEOR	ROUGE-L	CIDEr
Hard-Attention [48]	25.0	23.04	-	-
Review [51]	29.0	23.7	-	88.6
ATT [52]	30.4	24.3	-	-
CNN-RNN	31.9	25.3	53.55	99.1
CNN-RNN-RL	33.0	25.0	54.5	103.0

sequence model, but the sentences are influent sometimes. Aside from this, we find that there is a large semantic leap between two lines in PPG poems sometimes, and that the sentences are generally rigid. We infer that this may be caused by the fixed topic for each line.

4.5.5 Comparison between Poetry Keywords and Caption Words. To justify that the retrieval scheme is reasonable, we conduct a comparison between the number of poetry keywords and caption words. We randomly select 5,000 translated captions and do word segmentation. Then, we calculate the number of caption word segments (eliminating stop words) and the number of those exist in the poetry keyword dataset. The comparison can be seen in Figure 6. There are 27,568 different words in the poetry keywords dataset. 2,267 different caption words can be found in the poetry keyword dataset, which is 56.8% of 3,994 caption words. We also conduct TextRank algorithm on these 5,000 captions and extract maximal four caption-context keywords per caption. We find that 61.8% (1,538 of 2,489 caption keywords) caption-context keywords can be found in the poetry keyword dataset. In our method, maximal four visual keywords are required. Even though all the segmented words cannot be found in the poetry keyword dataset, we find the closest neighbour according to word2vec distance.

4.5.6 Image-captioning Performance. As image captioning is an important step in our approach, we present the performance for captioning as well.

Our captioning model is a variation of NIC [38]. We implement it with extra training tricks and discard the fine-tuning process. We show the performance of the model trained by cross-entropy loss and the model trained with RL algorithm, respectively. Performance of a bunch of existing captioning methods are listed as well. Hard-Attention [48] is the model using attention mechanism. Review network [51] is the one involving a review network between the encoder and the decoder. ATT [52] uses image attributes as well as CNN features to generate captions.



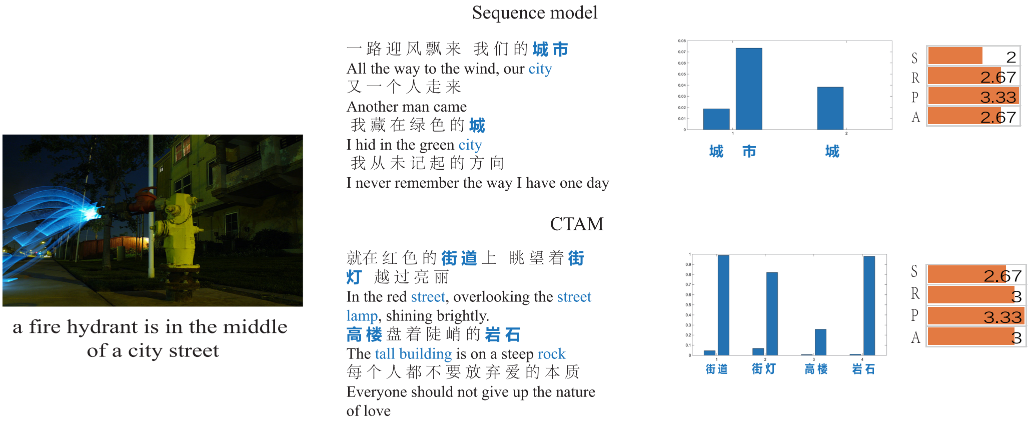
a bunk bed with a wooden frame in a bedroom
KWs: 床 bed 木 wooden
卧室 bedroom

a traffic light with a red light on it
KWs: 亮 light on 红灯red light
红绿灯 traffic light

a flock of birds flying over a beach
KWs: 鸟 bird 飞过 flying
海滩 beach

a dog laying on the sidewalk next to a bike
KWs: 躺 laying 狗 dog
旁边 next to 自行车 bike

Fig. 7. Visualization of image captions and related keywords (KWs in the figure).



a fire hydrant is in the middle of a city street

Sequence model

一路迎风飘来 我们的城市
All the way to the wind, our city
又一个人走来
Another man came
我藏在绿色的城
I hid in the green city
我从未记起的方向
I never remember the way I have one day

CTAM

就在红色的街道上 眺望着街灯 越过亮丽
In the red street, overlooking the street lamp, shining brightly.
高楼盘着陡峭的岩石
The tall building is on a steep rock
每个人都不要放弃爱的本质
Everyone should not give up the nature of love

Fig. 8. Visualization of generated results of Sequence model and the CTAM. We manually highlight relative words in poems, and present their generation probabilities in blue bars. Human evaluation results on Semantic consistency (S), Readability (R), Poeticness/Aesthetics (P), and the Average score (A) are shown in orange stripes.

The captioning performance is presented in Table 4. Based on Reference [38], captioning results are evaluated by four metrics: BLEU [29], METEOR [17], CIDEr [37], and ROUGE-L [20]. BLEU analyzes the co-occurrences of n-grams between the candidate and reference sentences. METEOR is calculated by generating a 1:1 alignment between the words in the candidate and reference sentence, which is highly correlated with human judgments in settings with a low number of references. ROUGE is designed for text summarization evaluation. The CIDEr metric performs a Term Frequency Inverse Document Frequency (TF-IDF) weighting for each n-gram. As it measures consensus in image captions, CIDEr is used in the validation process to select the best model, and it is used as a reward function in the RL training.

4.6 Qualitative Evaluation

In Figure 7, we present images with captions and related keywords. It can be seen that the selected keywords contain rich visual information and can cover the major parts of the images. Thus, we believe that our visual semantic vector is useful to convey visual information, and that it's effective for semantic consistency between images and poetries.

A pair of generation examples are shown in Figure 8. We present the comparison between the sequence model and the CTAM. We manually highlight the generated image-relative words in the poems, and we present their probabilities in blue bars. It can be seen that the CTAM

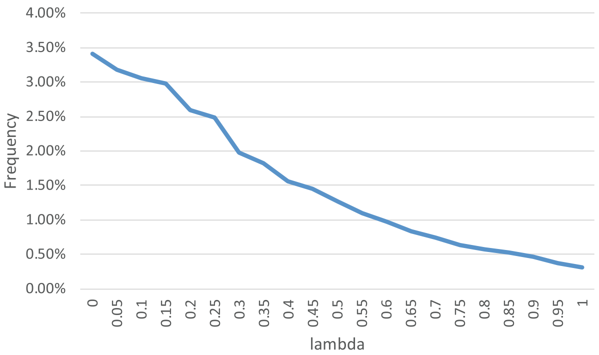


Fig. 9. Frequency of character “I” with respect to variant hyperparameter λ.

generation contains more related words compared with the sequence model generation. The human evaluations are illustrated as well. For the presented poems, the CTAM outperforms the sequence model across all the criteria. The human evaluation complementarily demonstrates the effect of our proposed model.

4.7 Effects of Hyperparameter λ

We test the effects of hyperparameter λ in Equation (14). We constrain the top 100 characters in the training corpus, and vary λ from 0 to 1. We select 200 keywords combination and test the frequency of character “I” among the whole generation data. The experiment results in Figure 9 show that the frequency decreases accordingly with the increase of λ. Interestingly, “I” is still among the top frequency characters in the generation, but the occurrence decreases.

5 CONCLUSION

In this article, we have proposed a Constrained Topic-aware Model (CTAM) to automatically create modern Chinese poems from images. This is a pioneer study on modern Chinese poetry creation from visual information, and the framework is originally designed. Among the three technical modules, a topic semantic vector is constructed based on the image-captioning model, which can ensure the semantic consistency between poems and images. The topic-aware generator provides a unique strategy to involve context information, which can avoid the topic drift problem. Decoding algorithms/strategies are worth investigating and they are important studies in the recurrent language modeling. Our proposed decoding scheme is straightforward and innovative, where it can penalize frequent words for the imbalanced dataset. Experiment results demonstrate that our approach can achieve promising performance, especially on readability and semantic consistency with the images. In the future, we plan to introduce the innovative generator into other natural language processing tasks such as dialogue generation.

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR'15).*

[2] Lalit Bahl, Peter Brown, Peter De Souza, and Robert Mercer. 1986. Maximum mutual information estimation of hidden Markov model parameters for speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'86)*, Vol. 11. IEEE, 49–52.

[3] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

[4] Simon Colton, Jacob Goodwin, and Tony Veale. 2012. Full-FACE poetry generation. In *Proceedings of the International Conference on Computational Creativity (ICCC'12)*. 95–102.

[5] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. 2018. Paying more attention to saliency: Image captioning with saliency and context attention. *ACM Trans. Multimedia Comput. Commun. Appl.* 14, 2 (2018), 48.

[6] Ali Farhadi, Mohsen Hejrati, Mohammad Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *Proceedings of the European Conference on Computer Vision (ECCV'10)*. 15–29.

[7] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017. Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3137–3146.

[8] Marjan Ghazvininejad, Xing Shi, Yejin Choi, and Kevin Knight. 2016. Generating topical poetry. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 1183–1191.

[9] Chen He and Haifeng Hu. 2019. Image captioning with visual-semantic double attention. *ACM Trans. Multimedia Comput. Commun. Appl.* 15, 1 (2019), 26.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[11] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.* 9, 8 (1997), 1735–1780.

[12] Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artific. Intell. Res.* 47 (2013), 853–899.

[13] Jack Hopkins and Douwe Kiela. 2017. Automatically generating rhythmic verse with neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 168–178.

[14] Nal Kalchbrenner, Ivo Danihelka, and Alex Graves. 2016. Grid long short-term memory. In *Proceedings of the International Conference on Learning Representations (ICLR'16)*.

[15] Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR'15)*.

[16] Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2011. Baby talk: Understanding and generating image descriptions. In *Proceedings of the 24th Conference on Computer Vision and Pattern Recognition (CVPR'11)*. Citeseer.

[17] Michael Denkowski Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL'14)*. 376.

[18] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT'16)*. 110–119.

[19] Siming Li, Girish Kulkarni, Tamara L. Berg, Alexander C. Berg, and Yejin Choi. 2011. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the 15th Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 220–228.

[20] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL'04)*, Vol. 8.

[21] Bei Liu, Jianlong Fu, Makoto P. Kato, and Masatoshi Yoshikawa. 2018. Beyond narrative description: Generating poetry from images by multi-adversarial training. In *Proceedings of the ACM Multimedia Conference on Multimedia Conference*. ACM, 783–791.

[22] Yu Liu, Jianlong Fu, Tao Mei, and Chang Wen Chen. 2017. Let your photos talk: Generating narrative paragraph for photo stream via bidirectional attention recurrent neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'17)*. 1445–1452.

[23] Alexander Patrick Mathews, Lexing Xie, and Xuming He. 2016. SentiCap: Generating image descriptions with sentiments. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'16)*. 3574–3580.

[24] Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

[25] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR'13)*.

[26] Yael Netzer, David Gabay, Yoav Goldberg, and Michael Elhadad. 2009. Gaiku: Generating haiku with word associations norms. In *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*. Association for Computational Linguistics, 32–39.

[27] Hugo Gonçalo Oliveira. 2012. PoeTryMe: A versatile platform for poetry generation. *Comput. Creat. Concept Invent. Gen. Intell.* 1 (2012), 21.

[28] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems.* MIT Press, 1143–1151.

[29] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics.* Association for Computational Linguistics, 311–318.

[30] Shengsheng Qian, Tianzhu Zhang, and Changsheng Xu. 2016. Multi-modal multi-view topic-opinion mining for social event analysis. In *Proceedings of the 24th ACM International Conference on Multimedia.* 2–11.

[31] Shengsheng Qian, Tianzhu Zhang, Changsheng Xu, and Jie Shao. 2015. Multi-modal event topic model for social event analysis. *IEEE Trans. Multimedia* 18, 2 (2015), 233–246.

[32] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR'16).*

[33] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'17)*, Vol. 1. 3.

[34] Bob L. Sturm, Joao Felipe Santos, Oded Ben-Tal, and Iryna Korshunova. 2016. Music transcription modelling and composition using deep learning. *arXiv preprint arXiv:1604.08723.*

[35] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems.* MIT Press, 3104–3112.

[36] Naoko Tosa, Hideto Obara, and Michihiko Minoh. 2008. Hitch haiku: An interactive supporting system for composing haiku poem. In *Proceedings of the International Conference on Entertainment Computing.* Springer, 209–216.

[37] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 4566–4575.

[38] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 3156–3164.

[39] Anqi Wang, Haifeng Hu, and Liang Yang. 2018. Image captioning with affective guiding and selective attention. *ACM Trans. Multimedia Comput. Commun. Appl.* 14, 3 (2018), 73.

[40] Cheng Wang, Haojin Yang, and Christoph Meinel. 2018. Image captioning with deep bidirectional lstms and multitask learning. *ACM Trans. Multimedia Comput. Commun. Appl.* 14, 2s (2018), 40.

[41] Qixin Wang, Tianyi Luo, Dong Wang, and Chao Xing. 2016. Chinese song iambics generation with neural attention-based model. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'16).*

[42] Zhe Wang, Wei He, Hua Wu, Haiyang Wu, Wei Li, Haifeng Wang, and Enhong Chen. 2016. Chinese poetry generation with planning based neural network. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers (COLING'16).* 1051–1060.

[43] Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. In *Reinforcement Learning.* Springer, 5–32.

[44] Jie Wu, Haifeng Hu, and Yi Wu. 2018. Image captioning via semantic guidance attention and consensus selection strategy. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14, 4 (2018), 87.

[45] Lingxiang Wu, Min Xu, Jinqiao Wang, and Stuart Perry. 2020. Recall what you see continually using GridLSTM in image captioning. *IEEE Trans. Multimedia* 22, 3 (2020), 808–818.

[46] Xiaofeng Wu, Naoko Tosa, and Ryohei Nakatsu. 2009. New hitch haiku: An interactive renku poem composition supporting tool applied for sightseeing navigation system. In *Proceedings of the International Conference on Entertainment Computing.* Springer, 191–196.

[47] Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'17)*, Vol. 17. 3351–3357.

[48] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning.* 2048–2057.

[49] Linli Xu, Liang Jiang, Chuan Qin, Zhe Wang, and Dongfang Du. 2018. How images inspire poems: Generating classical Chinese poetry from images with memory networks. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI'18).*

[50] Shijie Yang, Liang Li, Shuhui Wang, Weigang Zhang, Qingming Huang, and Qi Tian. 2019. SkeletonNet: A hybrid network with a skeleton-embedding process for multi-view image representation learning. *IEEE Trans. Multimedia* 21, 11 (2019), 2916–2929.

[51] Zhilin Yang, Ye Yuan, Yuexin Wu, William W Cohen, and Ruslan R Salakhutdinov. 2016. Review networks for caption generation. In *Advances in Neural Information Processing Systems.* MIT Press, 2361–2369.

[52] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 4651–4659.

[53] Jiyuan Zhang, Yang Feng, Dong Wang, Yang Wang, Andrew Abel, Shiyue Zhang, and Andi Zhang. 2017. Flexible and creative Chinese poetry generation using neural memory. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL'17).* 1364–1373.

[54] Xingxing Zhang and Mirella Lapata. 2014. Chinese poetry generation with recurrent neural networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'14).* 670–680.

[55] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision.* 2223–2232.

Received April 2019; revised November 2019; accepted February 2020