

# Beyond Narrative Description: Generating Poetry from Images by Multi-Adversarial Training

Bei Liu\*  
Kyoto University  
liubei@dl.kuis.kyoto-u.ac.jp

Makoto P. Kato  
Kyoto University  
mpkato@acm.org

Jianlong Fu†  
Microsoft Research Asia  
jianf@microsoft.com

Masatoshi Yoshikawa  
Kyoto University  
yoshikawa@i.kyoto-u.ac.jp

## ABSTRACT

Automatic generation of natural language from images has attracted extensive attention. In this paper, we take one step further to investigate generation of poetic language (with multiple lines) to an image for automatic poetry creation. This task involves multiple challenges, including discovering poetic clues from the image (e.g., hope from green), and generating poems to satisfy both relevance to the image and poeticness in language level. To solve the above challenges, we formulate the task of poem generation into two correlated sub-tasks by multi-adversarial training via policy gradient, through which the cross-modal relevance and poetic language style can be ensured. To extract poetic clues from images, we propose to learn a deep coupled visual-poetic embedding, in which the poetic representation from objects, sentiments<sup>1</sup> and scenes in an image can be jointly learned. Two discriminative networks are further introduced to guide the poem generation, including a multi-modal discriminator and a poem-style discriminator. To facilitate the research, we have released two poem datasets by human annotators with two distinct properties: 1) the first human annotated image-to-poem pair dataset (with 8,292 pairs in total), and 2) to-date the largest public English poem corpus dataset (with 92,265 different poems in total). Extensive experiments are conducted with 8K images, among which 1.5K image are randomly picked for evaluation. Both objective and subjective evaluations show the superior performances against the state-of-the-art methods for poem generation from images.

\*This work was conducted when Bei Liu was a research intern at Microsoft Research.

†Corresponding author

<sup>1</sup>We consider both adjectives and verbs that can express emotions and feelings as sentiment words in this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

MM '18, October 22–26, 2018, Seoul, Republic of Korea

© 2018 Copyright held by the owner/author(s). Publication rights licensed to the Association for Computing Machinery.

ACM ISBN 978-1-4503-5665-7/18/10...\$15.00

<https://doi.org/10.1145/3240508.3240587>



**Figure 1:** Example of human written description and poem of the same image. We can see a significant difference from words of the same color in these two forms. Instead of describing facts in the image, poem tends to capture deeper meaning and poetic symbols from objects, scenes and sentiments from the image (such as knight from *falcon*, hunting and fight from *eating*, and waiting from *standing*).

Turing test carried out with over 500 human subjects, among which 30 evaluators are poetry experts, demonstrates the effectiveness of our approach.

## CCS CONCEPTS

• Computing methodologies → Natural language generation; Image representations; Sequential decision making; Adversarial learning;

## KEYWORDS

Image, Poetry Generation, Adversarial Training

### ACM Reference Format:

Bei Liu, Jianlong Fu, Makoto P. Kato, and Masatoshi Yoshikawa. 2018. Beyond Narrative Description: Generating Poetry from Images by Multi-Adversarial Training. In *2018 ACM Multimedia Conference (MM '18)*, October 22–26, 2018, Seoul, Republic of Korea. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3240508.3240587>

## 1 INTRODUCTION

Researches that involve both vision and languages have attracted great attentions recently as we can witness from the bursting works on image descriptions like image caption and paragraph [1, 4, 17, 29]. Image descriptions aim to generate sentence(s) to describe facts from images in human-level languages. In this paper, we take one step further to tackle a more cognitive task: generation of poetic language to an image for the purpose of poetry creation, which has attracted tremendous interest in both research and industry fields.

In natural language processing field, poem generation related problems have been studied. In [12, 35], the authors mainly focused on the quality of style and rhythm. In [8, 35, 41], these works have taken one more step to generate poems from topics. Image inspired Chinese quatrain generation is proposed in [33]. In the industrial field, Facebook has proposed to generate English rhythmic poetry with neural networks [12], and Microsoft has developed a system called XiaoIce, in which poem generation is one of the most important features. Nevertheless, generating poems from images in an end-to-end fashion remains a new topic with grand challenges.

Compared with image captioning and paragraphing that focus on generating descriptive sentences about an image, generation of poetic language is a more challenging problem. There is a larger gap between visual representations and poetic symbols that can be inspired from images and facilitate better generation of poems. For example, “man” detected in image captioning can further indicate “hope” with “bright sunshine” and “opening arm”, or “loneliness” with “empty chairs” and “dark” background in poem creation. Fig. (1) shows a concrete example of the differences between descriptions and poems for the same image.

In particular, to generate a poem from an image, we are facing with the following three challenges. First of all, it is a cross-modality problem compared with poem generation from topics. An intuitive way for poem generation from images is to first extract keywords or captions from images and then consider them as seeds for poem generation as what poem generation from topics do. However, keywords or captions will miss a lot of information in images, not to mention the poetic clues that are important for poem generation [8, 41]. Secondly, compared with image captioning and image paragraphing, poem generation from images is a more subjective task, which means an image can be relevant to several poems from various aspects while image captioning/paragraphing is more about describing facts in the images and results in similar sentences. Thirdly, the form and style of poem sentences is different from that of narrative sentences. In this research, we mainly focus on *free verse* which is an open form of poetry. Although we do not require meter, rhyme or other traditional poetic techniques, it remains some sense of poetic structures and poetic style language in poems. We define this quality of poem as *poeticness* in this research. For example, length of poems are usually not very long, specific words are preferred in poems compared with image descriptions, and sentences in one poem should be consistent to one topic.

To address the above challenges, we collect two poem datasets by human annotators, and propose poetry creation by integrating retrieval and generation techniques in one system. Specifically, to better learn poetic clues from images for poem generation, we first learn a deep coupled visual-poetic embedding model with CNN features of images, and skip-thought vector features [16] of poems from a multi-modal poem dataset (namely “MultiM-Poem”) that consists of thousands of image-poem pairs. This embedding model is then used to retrieve relevant and diverse poems from a larger uni-modal poem corpus (namely “UniM-Poem”) for images.

Images with these retrieved poems and MultiM-Poem together construct an enlarged image-poem pair dataset (namely “MultiM-Poem (Ex)”). We further propose to leverage the state-of-art sequential learning techniques for training an end-to-end image to poem model on the MultiM-Poem (Ex) dataset. Such a framework ensures substantial poetic clues, that are significant for poem generation, could be discovered and modeled from those extended pairs.

To avoid exposure bias problems caused by long length of long sequence (all poem lines together) and the problem that there is no specific loss available to score a generated poem, we propose to use a recurrent neural network (RNN) for poem generation with multi-adversarial training and further optimize it by policy gradient. Two discriminative networks are used to provide rewards in terms of the generated poem’s relevance to the given image and poeticness of the generated poem. We conduct experiments on MultiM-Poem, UniM-Poem and MultiM-Poem (Ex) to generate poems to images. The generated poems are evaluated in both objective and subjective ways. We define automatic evaluation metrics concerning relevance, novelty and translative consistence and conducted user studies about relevance, coherence and imaginativeness of generated poems to compare our model with baseline methods. The contributions in this research are concluded as follows:

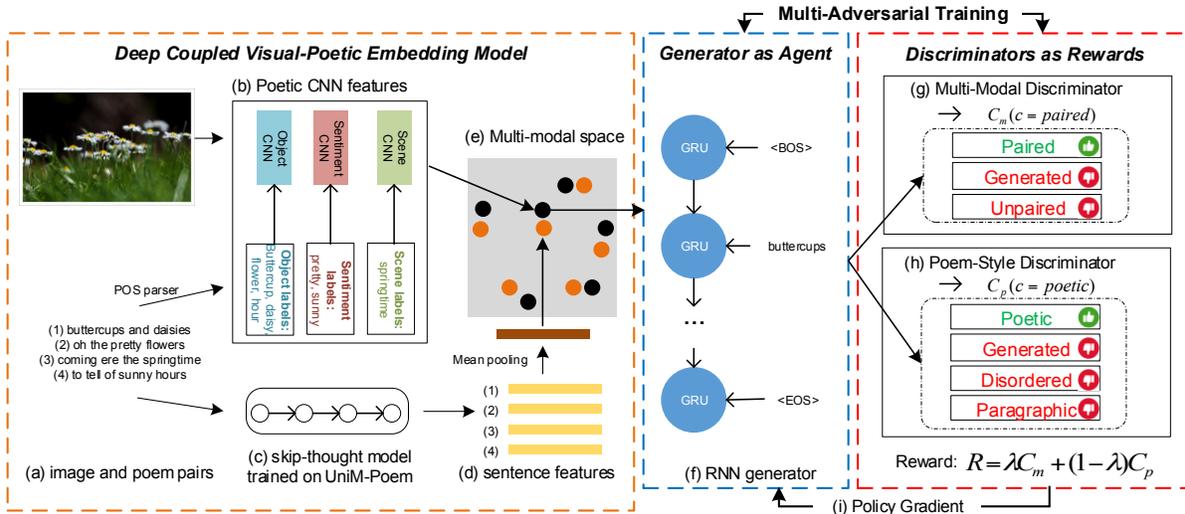
- We propose to generate poems (English free verse) from images in an end-to-end fashion. To the best of our knowledge, this is the first attempt to study the image-inspired English poem generation problem in a holistic framework, which enables a machine to approach human capability in cognition tasks.
- We incorporate a deep coupled visual-poetic embedding model and a RNN-based generator for joint learning, in which two discriminators provide rewards for measuring cross-modality relevance and poeticness by multi-adversarial training.
- We collect the first paired dataset of image and poem annotated by human annotators, and the largest public poem corpus dataset. Extensive experiments demonstrate the effectiveness of our approach compared with several baselines by using both objective and subjective evaluation metrics, including a Turing test from more than 500 human subjects. To better promote the research in poetry generation from images, we have released these datasets and our code on Github<sup>2</sup>.

## 2 RELATED WORK

### 2.1 Poetry Generation

Traditional approaches for poetry generation include template and grammar-based method [20–22], generative summarization under constrained optimization [35] and statistical machine translation model [11, 13]. By applying deep learning approaches recent years, researches about poetry generation has entered a new stage. Recurrent neural network

<sup>2</sup><https://github.com/bei21/img2poem>



**Figure 2: The framework of poetry generation with multi-adversarial training.** A deep coupled visual-poetic model (e) is trained by human annotated image-poem pairs (a). The image features (b) are poetic multi-CNN features obtained by fine-tuning CNNs with the extracted poetic symbols (e.g., objects, scenes and sentiments) by a POS parser [28] from poems. The sentence features (d) of poems are extracted from a skip-thought model (c) trained on the largest public poem corpus (UniM-Poem). A RNN-based sentence generator (f) is trained as agent and two discriminators considering multi-modal (g) and poem-style (h) critics of a generated poem to a given image provide rewards to policy gradient (i). POS parser extracts Part-Of-Speech words from poems.

is widely used to generate poems that can even confuse readers from telling them from poems written by human poets [8, 9, 12, 37, 41]. Previous works of poem generation mainly focus on style and rhythmic qualities of poems [12, 35], while recent studies introduce topic as a condition for poem generation [8, 9, 35, 41]. For a poem, topic is still a rather abstract concept without specific scenarios. Inspired by the fact that many poems were created in a conditioned scenario, we take one step further to tackle the problem of generating poems inspired by a visual scenario. Compared with previous researches, our work is facing with more challenges, especially in terms of multi-modal problems.

## 2.2 Image Description

Image captioning is first regarded as a retrieval problem which aims to search captions from dataset for a given image [5, 14] and hence cannot provide accurate and proper descriptions for all images. To overcome this problem, methods like template filling [18] and paradigm for integrating convolutional neural network (CNN) and recurrent neural network (RNN) [2, 29, 36, 38] are proposed to generate readable human-level sentences. Recently, generative adversarial network (GAN) is applied to generate captions based on different problem settings [1, 39]. Similarly to image captioning, image paragraphing is going the similar way. Recent researches about image paragraphing mainly focus on region detection and hierarchical structure for generated sentences [17, 19, 24]. However, as we have addressed, image captioning and paragraphing aim to generate descriptive sentences to tell the facts in images, while poem generation is tackling an advanced form of linguistic form which requires poeticness and language style constrains.

## 3 APPROACH

In this research, we aim to generate poems from images so that the generated poems are relevant to input images and satisfy poeticness. For this purpose, we cast our problem in a multi-adversarial procedure [10] and further optimize it with a policy gradient [32, 40]. A CNN-RNN generative model acts as an *agent*. The parameters of this agent define a *policy* whose execution will decide which word to be picked as an *action*. When the agent has picked all words in a poem, it observes a *reward*. We define two discriminative networks to serve as rewards concerning whether the generated poem is a paired one with the input image and whether the generated poem is poetic. The goal of our poem generation model is to generate a sequence of words as a poem for an image to maximize the expected end reward. This policy-gradient method has shown significant effectiveness to many tasks without non-differentiable metrics [1, 25, 39].

As shown in Fig. (2), the framework consists of several parts: (1) a deep coupled visual-poetic embedding model to learn poetic representations from images, and (2) a multi-adversarial training procedure optimized by policy gradient. A RNN based generator serves as agent, and two discriminative networks provide rewards to the policy gradient.

### 3.1 Deep Coupled Visual-Poetic Embedding

The goal of visual-poetic embedding model [6, 15] is to learn an embedding space where points of different modality, e.g. images and sentences, can be projected to. In a similar way to image captioning problem, we assume that a pair of image and poem shares similar poetic semantics which makes the embedding space learnable. By embedding both images and poems to the same feature space, we can directly compute

the relevance between a poem and an image by poetic vector representations of them. Moreover, the embedding feature can be further utilized to initialize a optimized representation of poetic clues for poem generation.

The structure of our deep coupled visual-poetic embedding model is shown in left part of Fig. (2). For the input of images, we leverage three deep convolutional neural networks (CNNs) concerning three aspects that indicate important poetic clues from images inspired from fine-grained problems [7], namely object ( $\mathbf{v}_1$ ), scene ( $\mathbf{v}_2$ ) and sentiment ( $\mathbf{v}_3$ ), after conducting a prior user study about important factors for poem creation from images. We observed that concepts in poems are often imaginative and poetic while concepts in the classification datasets we use to train our CNN models are concrete and common. To narrow the semantic gap between the visual representation of images and the textual representation of poems, we propose to fine-tune these three networks with MultiM-Poem dataset. Specifically, frequent used keywords about object, sentiment and scenes in the poems are picked as label vocabulary, and then we build three multi-label datasets based on MultiM-Poem dataset for object, sentiment and scenes detection respectively. Once the multi-label datasets are built, we fine-tune the pre-trained CNN models on the three datasets independently, which is optimized by sigmoid cross entropy loss as shown in Eq. (1). After that, we adopt the  $D$ -dimension deep features for each aspect from the penultimate fully-connected layer of the CNN models, and get a concatenated  $N$ -dimension ( $N = D \times 3$ ) feature vector  $\mathbf{v} \in \mathbb{R}^N$  as input of visual-poetic embedding for each image:

$$\begin{aligned} \text{loss} &= \frac{-1}{N} \sum_{n=1}^N (t_n \log p_n + (1 - t_n) \log(1 - p_n)), & (1) \\ \mathbf{v}_1 &= f_{\text{Object}}(I), \quad \mathbf{v}_2 = f_{\text{Scene}}(I), & (2) \\ \mathbf{v}_3 &= f_{\text{Sentiment}}(I), \quad \mathbf{v} = (\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3). \end{aligned}$$

The output of visual-poetic embedding vector  $\mathbf{x}$  is a  $K$ -dimension vector representing the image embedding with linear mapping from image features:

$$\mathbf{x} = \mathbf{W}_v \cdot \mathbf{v} + \mathbf{b}_v \in \mathbb{R}^K, \quad (3)$$

where  $\mathbf{W}_v \in \mathbb{R}^{K \times N}$  is the image embedding matrix and  $\mathbf{b}_v \in \mathbb{R}^K$  is the image bias vector. Meanwhile, representation feature vector of a poem is computed by skip-thought vectors[16], which is a popular unsupervised method to learn sentence embedding. We train skip-thought model on unpaired UniM-Poem dataset and use it to provide a better sentence representation for poem sentences. Mean value of all sentences' combined skip-thought features (unidirectional and bidirectional) is denoted by  $\mathbf{t} \in \mathbb{R}^M$  where  $M$  is the combined dimension. Similar to image embedding, the poem embedding is denoted as:

$$\mathbf{m} = \mathbf{W}_t \cdot \mathbf{t} + \mathbf{b}_t \in \mathbb{R}^K, \quad (4)$$

where  $\mathbf{W}_t \in \mathbb{R}^{K \times M}$  for the poem embedding matrix and  $\mathbf{b}_t \in \mathbb{R}^K$  for the poem bias vector. Finally, the image and poem are embedded together by minimizing a pairwise ranking loss with dot-product similarity:

$$\begin{aligned} L &= \sum_{\mathbf{x}} \sum_k \max(0, \alpha - \mathbf{x} \cdot \mathbf{m} + \mathbf{x} \cdot \mathbf{m}_k) \\ &+ \sum_{\mathbf{m}} \sum_k \max(0, \alpha - \mathbf{m} \cdot \mathbf{x} + \mathbf{m} \cdot \mathbf{x}_k), \end{aligned} \quad (5)$$

where  $\mathbf{m}_k$  is a contrastive (irrelevant unpaired) poem for image embedding  $\mathbf{x}$ , and vice-versa with  $\mathbf{x}_k$ .  $\alpha$  denotes the contrastive margin. As a result, the model we trained will produce higher dot-product similarity between embedding features of image-poem pairs than similarity between randomly generated pairs.

### 3.2 Poem Generator as an Agent

A conventional CNN-RNN model for image captioning is used to serve as an agent. Instead of using hierarchical methods that are used recently in generating multiple sentences [17], we use a non-hierarchical recurrent model by treating the end-of-sentence token as a word in the vocabulary. The reason is that 1) poems often consist of fewer words compared with paragraphs; 2) there is lower consistent hierarchy between sentences of poems, which makes the hierarchy much more difficult to learn. We also conduct experiment with hierarchical recurrent language model as a baseline and we will show the result in the experiment part.

The generative model includes CNNs for image encoder and a RNN for poem decoder. The reason of using RNN instead of CNN for languages is that it can better encode the structure-dependent semantics of the long sentences which are widely observed in poems. In this research, we apply Gated Recurrent Units (GRUs) [3] for poem decoder for its simple structure and robustness to overfitting problem on less training data. We use image-embedding features learned by the deep coupled visual-poetic embedding model explained in Section 3.1 as input of image encoder. Suppose  $\theta$  is the parameters of the model. Traditionally, our target is to learn  $\theta$  by maximizing the likelihood of the observed sentence  $\mathbf{y} = y_{1:T} \in \mathbb{Y}^*$  where  $T$  is the maximum length of generated sentence (including  $\langle \text{BOS} \rangle$  for start of sentence,  $\langle \text{EOS} \rangle$  for end of sentence and line breaks) and  $\mathbb{Y}^*$  denotes a space of all sequences of selected words.

Let  $r(y_{1:t})$  denote the reward achieved at time  $t$  and  $R(y_{1:T})$  is the cumulative reward, namely  $R(y_{k:T}) = \sum_{t=k}^T r(y_{1:t})$ . Let  $p_\theta(y_t | y_{1:(t-1)})$  be a parametric conditional probability of selecting  $y_t$  at time step  $t$  given all the previous words  $y_{1:(t-1)}$ .  $p_\theta$  is defined as a parametric function of policy  $\theta$ . The reward of policy gradient in each batch can be computed as the sum over all sequences of valid actions as the expected future reward. To iterate over sequences of all possible actions is exponential, but we can further write it in expectation so that it can be approximated with an unbiased estimator:

$$J(\theta) = \sum_{y_{1:T} \in \mathbb{Y}^*} p_\theta(y_{1:T}) R(y_{1:T}) = \mathbb{E}_{y_{1:T} \sim p_\theta} \sum_{t=1}^T r(y_{1:t}). \quad (6)$$

We aim to maximize  $J(\theta)$  by following its gradient:

$$\nabla_\theta J(\theta) = \mathbb{E}_{y_{1:T} \sim p_\theta} \left[ \sum_{t=1}^T \nabla_\theta \log p_\theta(y_{1:t-1}) \right] \sum_{t=1}^T r(y_{1:t}). \quad (7)$$

In practice the expected gradient can be approximated using a Monte-Carlo sample by sequentially sample each  $y_t$  from

the model distribution  $p_\theta(y_t|y_{1:(t-1)})$  for  $t$  from 1 to  $T$ . As discussed in [25], a baseline  $b$  can be introduced to reduce the variance of the gradient estimate without changing the expected gradient. Thus, the expected gradient with a single sample is approximated as follow:

$$\nabla_\theta J(\theta) \approx \sum_{t=1}^T \nabla_\theta \log p_\theta(y_{1:t-1}) \sum_{t=1}^T (r(y_{1:t}) - b_t). \quad (8)$$

### 3.3 Discriminators as Rewards

A good poem for an image has to satisfy at least two criteria: the poem (1) is relevant to the image, and (2) has some sense of poeictiness concerning proper length, poem’s language style and consistence between sentences. Based on these two requirements, we propose two discriminative networks to guide the generated poem: multi-modal discriminator and poem-style discriminator.

**Multi-Modal Discriminator.** Multi-modal discriminator ( $D_m$ ) is used to guide the generated poem  $\mathbf{y}$  related to corresponding image  $\mathbf{x}$ . It is trained to classify a poem into three classes: *paired* as positive examples, *unpaired* and *generated* as negative examples. *Paired* includes ground-truth paired poems for the input images. *Unpaired* poems are randomly sampled from unpaired poems of the input images in training data.  $D_m$  includes a multi-modal encoder, modality fusion layer and a classifier with softmax function:

$$\mathbf{c} = \text{GRU}_\rho(\mathbf{y}), \quad (9)$$

$$f = \tanh(W_x \cdot \mathbf{x} + b_x) \odot \tanh(W_c \cdot \mathbf{c} + b_c), \quad (10)$$

$$C_m = \text{softmax}(W_m \cdot f + b_m), \quad (11)$$

where  $\rho$ ,  $W_x$ ,  $b_x$ ,  $W_c$ ,  $b_c$ ,  $W_m$ ,  $b_m$  are parameters to be learned,  $\odot$  is element-wise multiplication and  $C_m$  denotes the probabilities over three classes of the multi-modal discriminator. We utilize GRU-based sentence encoder for discriminator training. Eq. (11) provides way to generate the probability of  $(\mathbf{x}, \mathbf{y})$  classified into each class as denoted by  $C_m(c|\mathbf{x}, \mathbf{y})$  where  $c \in \{\textit{paired}, \textit{unpaired}, \textit{generated}\}$ .

**Poem-Style Discriminator.** In contrast with most poem generation researches that emphasize on meter, rhyme or other traditional poetic techniques, we focus on *free verse* which is an open form of poetry. Even though, we require our generated poems have the quality of *poeticness* as we define in Section 1. Without making specific templates or rules for poems, we propose a poem-style discriminator ( $D_p$ ) to guide generated poems towards human written poems. In  $D_p$ , generated poems will be classified into four classes: *poetic*, *disordered*, *paragraphic* and *generated*.

Class *poetic* is addressed as positive example of poems that satisfy poeticness. The other three classes are all regarded as negative examples. Class *disordered* concerns about the inner structure and coherence between sentences of poems and *paragraphic* class uses paragraph sentences as negative examples. In  $D_p$ , we use UniM-Poem as positive *poetic* samples. To construct disordered poems, we first construct a poem sentence pool by splitting all poems in UniM-Poem. Examples of class *disordered* are poems that we reconstruct

by sentences randomly picked up with a reasonable line numbers from poem sentence pool. Paragraph dataset provided by [17] is used as *paragraph* examples.

A completed generated poem  $\mathbf{y}$  is encoded by GRU and parsed to a fully connected layer, and the probability of falling into four classes is computed by a softmax function. Formula of this procedure is as follow:

$$C_p = \text{softmax}(W_p \cdot \text{GRU}_\eta(\mathbf{y}) + b_p), \quad (12)$$

where  $\eta$ ,  $W_p$ ,  $b_p$  are parameters to be learned. The probability of classifying generated poem  $\mathbf{y}$  to a class  $c$  is formulated as  $C_p(c|\mathbf{y})$  where  $c \in \{\textit{poetic}, \textit{disordered}, \textit{paragraphic}, \textit{generated}\}$ .

**Reward Function.** We define the reward function for policy gradient as a linear combination of probability of classifying generated poem  $\mathbf{y}$  for an input image  $\mathbf{x}$  to the positive class (*paired* for multi-modal discriminator  $D_m$  and *poetic* for poem-style discriminator  $D_p$ ) weighted by tradeoff parameter  $\lambda$ :

$$R(\mathbf{y}|\cdot) = \lambda C_m(c = \textit{paired}|\mathbf{x}, \mathbf{y}) + (1 - \lambda) C_p(c = \textit{poetic}|\mathbf{y}). \quad (13)$$

### 3.4 Multi-Adversarial Training

Adversarial training is a minimax game between a generator  $G$  and a discriminator  $D$  with value function  $V(G, D)$ :

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{\text{data}}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))]. \quad (14)$$

We propose to use multiple discriminators by reformulating  $G$ ’s objective as:

$$\min_G F(V(D_1, G), \dots, V(D_n, G)), \quad (15)$$

where we have  $n = 2$ , and  $F$  indicates linear combination of discriminators as shown in Eq. (13).

The generator aims to generate poems that have higher rewards for both discriminators so that they can fool the discriminators while the discriminators are trained to distinguish the generated poems from *paired* and *poetic* poems. The probabilities of classifying generated poem into positive classes in both discriminators are used as rewards to policy gradient as explained above.

Multiple discriminators (two in this work) are trained by providing positive examples from the real data (*paired* poems in  $D_m$  and poem corpus in  $D_p$ ) and negative examples from poems generated from the generator as well as other negative forms of real data (*unpaired* poems in  $D_m$ , *paragraphs* and *disordered* poems in  $D_p$ ). Meanwhile, by employing a policy gradient and Monte Carlo sampling, the generator is updated based on the expected rewards from multiple discriminators. Since we have two discriminators, we apply a multi-adversarial training method that will train two discriminators in a parallel way.

## 4 EXPERIMENTS

### 4.1 Datasets

To facilitate the research of poetry generation from images, we collected two poem datasets, in which one consists of image and poem pairs, namely Multi-Modal Poem dataset (MultiM-Poem), and the other is a large poem corpus, namely

Name	#Poem	#Line/poem	#Word/line
MultiM-Poem	8,292	7.2	5.7
UniM-Poem	93,265	5.7	6.2
MultiM-Poem (Ex)	26,161	5.4	5.9

**Table 1: Detailed information about the three datasets. The first two datasets are collected by ourselves and the third one is extended by our embedding model.**

Uni-Modal Poem dataset (UniM-Poem). By using the embedding model we have trained, the image and poem pairs are extended by adding the nearest three neighbor poems from the poem corpus without redundancy, and an extended image and poem pair dataset is constructed and denoted as MultiM-Poem (Ex). The detailed information about these datasets is listed in Table 1.

For MultiM-Poem dataset, we first crawled 34,847 image-poem pairs in Flickr from groups that aim to use images illustrating poems written by human. Five human assessors majoring in English literature were further asked to evaluate these poems as relevant or irrelevant by judging whether the image can exactly inspire the poem in a pair by considering the associations of objects, sentiments and scenes. We filtered out pairs labeled as irrelevant and kept the remaining 8,292 pairs to construct the MultiM-Poem dataset.

UniM-Poem is crawled from several public online poetry websites, such as Poetry Foundation<sup>3</sup>, PoetrySoup<sup>4</sup>, best-poem.net and poets.org. To achieve robust model training, a poem pre-processing procedure is conducted to filter out those poems with too many lines ( $> 10$ ) or too fewer lines ( $< 3$ ). We also remove poems with strange characters, poems in languages other than English and duplicate poems.

## 4.2 Compared Methods

To investigate the effectiveness of the proposed methods, we compare with four baseline models with different settings. The models of show-and-tell [29] and SeqGAN [39] are selected due to their state-of-art results in image captioning. A competitive image paragraphing model is selected, as its strong capability for modeling diverse image content. Note that all the methods use MultiM-Poem (Ex) as the training dataset, and can generate multiple lines as poems. The detailed experiment settings are shown as follows:

**Show and tell (1CNN):** CNN-RNN model trained with only object CNN by VGG-16 .

**Show and tell (3CNNs):** CNN-RNN model trained with three CNN features by VGG-16.

**SeqGAN:** CNN-RNN model optimized with a discriminator to tell from generated poems and ground-truth poems. We use RNN for discriminator for fair comparison.

**Regions-Hierarchical:** Hierarchical paragraph generation model based on [17]. To better align with poem distribution, we restrict the maximum lines to be 10 and each line has up to 10 words in the experiment.

<sup>3</sup><https://www.poetryfoundation.org/>

<sup>4</sup><https://www.poetrysoup.com/>

**Our Model:** To demonstrate the effectiveness of the two discriminators, we train our model (Image to Poem with GAN, I2P-GAN) in four settings: pretrained model without discriminators (**I2P-GAN w/o discriminator**), with multi-modal discriminator only (**I2P-GAN w/  $D_m$** ), with poem-style discriminator only (**I2P-GAN w/  $D_p$** ) and with both discriminators (**I2P-GAN**).

## 4.3 Automatic Evaluation Metrics

Evaluation of poems is generally a difficult task and there are no established metrics in existing works, not to mention the new task of generating poems from images. To better address the performance of the generated poems, we propose to evaluate them in both automatic and manual way.

We propose to employ three metrics for automatic evaluation, e.g., BLEU, novelty and relevance. An overall score is computed by the three metrics after normalization.

**BLEU.** We use Bilingual Evaluation Understudy (BLEU) [23] score-based evaluation to examine how likely the generated poems can approximate towards the ground-truth ones following image captioning and paragraphing. It is also used in some poem generation works [35]. For each image, we only use the human written poems as ground-truth poems.

**Novelty.** By introducing discriminator  $D_p$ , the generator is supposed to introduce words or phrases from UniM-Poem dataset and results in words or phrases that are not very frequent in MultiM-Poem (Ex) dataset. We use *novelty* as proposed by [34] to measure the number of infrequent words or phrases observed in the generated poems. Two scales of N-gram are explored, e.g. bigram and trigram, as **Novelty-2** and **Novelty-3**. We first rank the n-grams that occur in the training dataset of MultiM-Poem (Ex) and take the top 2,000 as frequent ones. Novelty is computed as the proportion of n-grams that occur in training dataset except the frequent ones in the generated poem.

**Relevance.** Different from poem generation researches that have no or weak constrains to poem contents, we consider relevance of the generated poem to the given image as an important measurement in this research. However, unlike captions that concern more about facts about images, different poems can be relevant to the same image from various aspects. Thus, instead of computing relevance between generated poem and ground-truth poems, we define relevance between a poem and an image using our learned deep coupled visual-poetic embedding (VPE) model. After mapping the image and the poem to the same space through VPE, linearly scaled cosine similarity (0-1) is used to measure their relevance.

**Overall.** We compute an overall score based on the above three metrics. For each value  $a_i$  in all values of one metric  $a$ , we first linearly normalize it with following method:

$$a_i' = \frac{a_i - \min(a)}{\max(a) - \min(a)}. \quad (16)$$

After that, we get average values for BLEU (e.g. BLEU-1, BLEU-2 and BLEU-3) and novelty (e.g. Novelty-2 and Novelty-3). A final score is computed by averaging the normalized values, to ensure equal contribution of different metrics.

However, in such an open-ended task, there are no particularly suitable metrics that can perfectly evaluate the performance of generated poems. The automatic metrics we use can be used as a guidance to some extent. To better illustrate the performance of poems from human perception, we further conduct extensive user studies in the follows.

#### 4.4 Human Evaluation

We conducted human evaluation in Amazon Mechanical Turk. In particular, three types of tasks are assigned:

**Task1:** to explore the effectiveness of our deep coupled visual-poetic embedding model, annotators were requested to give a 0-10 scale score to a poem given an image considering their relevance in case of content, emotion and scene.

**Task2:** this task aims to compare the generated poems by different methods (four baseline methods and our four model settings) for one image on different aspects. Given an image, the annotators were asked to give ratings to a poem on a 0-10 scale with respect to four criteria: relevance (to the image), coherence (whether the poem is coherent across lines), imaginativeness (how much imaginative and creative the poem is for the given image) and overall impression.

**Task3:** Turing test was conducted by asking annotators to select human written poem from mixed human written and generated poems. Note that Turing test was implemented in two settings, i.e., with and without images as references.

For each task, we have randomly picked up 1K images and each task is assigned to three assessors. As poem is a form of literature, we also ask 30 annotators whose majors are related to English literature (among which ten annotations are English natives) as *expert users* to do the Turing test.

#### 4.5 Training Details

In the deep coupled visual-poetic embedding model, we use  $D = 4, 096$ -dimension “fc7” features for each CNN. Object features are extracted from VGG-16 [27] trained on ImageNet [26], scene features from Place205-VGGNet model [31], and sentiment features from sentiment model[30].

To better extract visual feature for poetic symbols, we first get nouns, verbs and adjectives with at least five frequency in UniM-Poem dataset. Then we manually picked adjectives and verbs for sentiment (including 328 labels), nouns for object (including 604 labels) and scenes (including 125 labels). As for poem features, we extract a combined skip-thought vector with  $M = 2, 048$ -dimension (in which each 1, 024-dimension represents for uni-direction and bi-direction, respectively) for each sentence, and finally we get poem features by mean pooling. And the margin  $\alpha$  is set to 0.2 based on empirical experiments in [15]. We randomly select 127 poems as unpaired poems for an image and used them as contrastive poems ( $\mathbf{m}_k$  and  $\mathbf{x}_k$  in Eq. (5)), and we re-sample them in each epoch. Before adversarial training, we pre-train a generator based on image captioning method [29] which can provide a better policy initialization for generator. We empirically set the tradeoff parameter  $\lambda = 0.8$  by conducting a comparable observation on automatic evaluation results from 0.1 to 0.9.

#### 4.6 Evaluations

	Ground-Truth	VPE w/o FT	VPE w/ FT
Relevance	7.22	5.82	6.32

**Table 2: Average score of relevance to images for three types of human written poems on 0-10 scale (0-irrelevant, 10-relevant). One-way ANOVA revealed that evaluation on these poems is statistically significant ( $F(2, 9) = 130.58, p < 1e - 10$ ).**

**Retrieved Poems.** We compare three kinds of poems considering their relevance to images: ground-truth poems, poems retrieved with VPE and image features before fine-tuning (VPE w/o FT), and poems retrieved with VPE and fine-tuned image features (VPE w/ FT). Table 2 shows a comparison on a scale of 0-10 (0 means irrelevant and 10 means the most relevant). We can see that by using the proposed visual-poetic embedding model, the retrieved poems can achieve a relevance score above the average score (i.e., the score of five). And image features fine-tuned with poetic symbols can improve the relevance significantly.

**Generated Poems.** Table 3 exhibits the automatic evaluation results of the proposed model with four settings, as well as the four baselines proposed in previous works. Comparing results of caption model with one CNN and three CNNs, we can see that multi-CNN can actually help to generate poems that are more relevant to images. Regions-Hierarchical model emphasizes more on the topic coherence between sentences while many human written poems will cover several topics or use different symbols for one topic. SeqGAN shows the advantage of applying adversarial training for poem generation compared with only caption models with only CNN-RNN while lacking of generating novel concepts in poems. Better performance of our pre-trained model with VPE than caption model demonstrates the effectiveness of VPE in extracting poetic features from images for better poem generation. We can see that our three models outperform in most of the metrics with each one performs better at one aspect. The model with only multi-modal discriminator (**I2P-GAN w/  $D_m$** ) will guide the model to generate poems towards ground-truth poems, thus it results in the highest BLEU scores that emphasize the similarity of n-grams in a translative way. Poem-style discriminator ( $D_p$ ) is designed to guide the generated poem to be more poetic in language style, and the highest novelty score of **I2P-GAN w/  $D_m$**  shows that  $D_p$  helps to provide more novel and imaginative words to the generated poem. Overall, **I2P-GAN** combines the advantages of both discriminators with a rational intermediate score regarding BLEU and novelty while still outperforms compared with other generation models. Moreover, our model with both discriminators can generate poems that have highest relevance on our embedding relevance metric.

Comparison of human evaluation results are shown in Table 4. Different from automatic evaluation results where Regions-Hierarchical performs not well, it gets a slightly better result than caption model for the reason that sentences all about the same topic tend to gain better impressions from users. Our three models outperform the other four

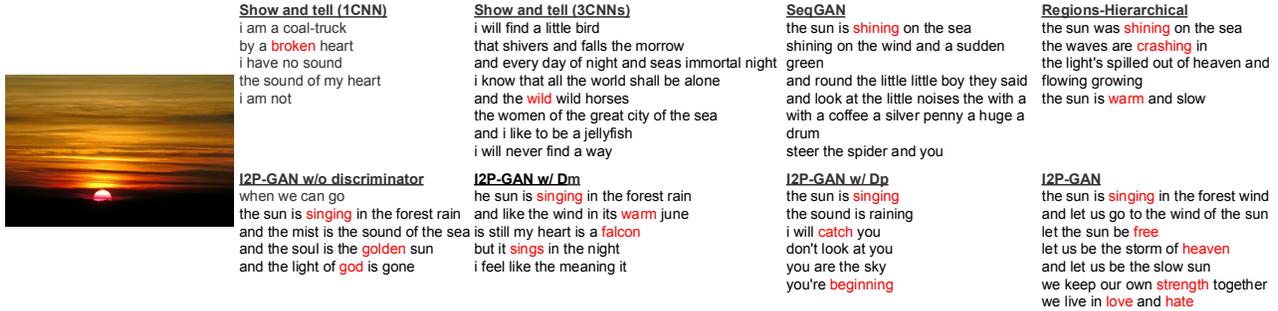


Figure 3: Example of poems generated by eight methods for an image. Words in read indicate poeticness.

Method	Relevance	Novelty-2	Novelty-3	BLEU-1	BLEU-2	BLEU-3	Overall
Show and Tell (1CNN)[29]	1.79	43.66	76.76	11.88	3.35	0.76	14.40
Show and Tell (3CNNs)[29]	1.91	48.09	81.37	12.64	3.34	0.8	34.34
SeqGAN[39]	2.03	47.52	82.32	13.40	3.72	0.76	44.95
Regions-Hierarchical[17]	1.81	46.75	79.90	11.64	2.5	0.67	8.01
I2P-GAN w/o discriminator	1.94	45.25	80.13	13.35	3.69	0.88	41.86
I2P-GAN w/ $D_m$	2.07	43.37	78.98	<b>15.15</b>	<b>4.13</b>	<b>1.02</b>	63.00
I2P-GAN w/ $D_p$	1.90	<b>60.66</b>	<b>89.74</b>	12.91	3.05	0.72	51.35
I2P-GAN	<b>2.25</b>	54.32	85.37	14.25	3.84	0.94	<b>77.23</b>

Table 3: Automatic evaluation. Note that BLEU scores are computed in comparison with human-annotated ground-truth poems (one poem for one image). Overall score is computed as an average of three metrics after normalization (Eq. (16)). All scores are reported as percentage (%).

Method	Rel	Col	Imag	Overall
Show and Tell (1CNN)[29]	6.31	6.52	6.57	6.67
Show and Tell (3CNNs)[29]	6.41	6.59	6.63	6.75
SeqGAN[39]	6.13	6.43	6.50	6.63
Regions-Hierarchical[17]	6.35	6.54	6.63	6.78
I2P-GAN w/o discriminator	6.44	6.64	6.77	6.85
I2P-GAN w/ $D_m$	6.59	6.83	6.94	7.06
I2P-GAN w/ $D_p$	6.53	6.75	6.80	6.93
I2P-GAN	<b>6.83</b>	<b>6.95</b>	<b>7.05</b>	<b>7.18</b>
Ground-Truth	7.10	7.26	7.23	7.37

Table 4: Human evaluation results of six methods on four criteria: relevance (Rel), coherence (Col), imaginativeness (Imag) and Overall. All criteria are evaluated on 0-10 scale (0-bad, 10-good).

baseline methods on all metrics. Two discriminators promote human-level comprehension towards poems compared with pre-trained model. The model with two discriminators has generated better poems from images in terms of relevance, coherence and imaginativeness. Fig. (3) shows one example of poems generated with three baselines and our methods for a given image. More examples generated by our approach can be referred in the supplementary material.

Data	Users	Ground-Truth	Generated
Poem w/ Image	AMT	0.51	0.49
	Expert	0.60	0.40
Poem w/o Image	AMT	0.55	0.45
	Expert	0.57	0.43

Table 5: Accuracy of Turing test on AMT users and expert users on poems with and without images.

**Turing Test.** For the Turing test of annotators in AMT, we have hired 548 workers with 10.9 tasks for each one on average. For experts, 15 people were asked to judge human written poems with images and another 15 annotators were

asked to do test with only poems. Each one is assigned with 20 images and in total we have 600 tasks conducted by expert users. Table 5 shows the probability of different poems being selected as human-written poems for a given image. As we can see, the generated poems have caused a competitive confusion to both ordinary annotators and experts though experts can figure out the accurate one better than ordinary people. One interesting observation comes from that experts are better at figuring out correct ones with images while AMT workers do better with only poems.

## 5 CONCLUSION

As the frontal work of poetry (English free verse) generation from images, we propose a novel approach to model the problem by incorporating deep coupled visual-poetic embedding model and RNN based adversarial training with multi-discriminators as rewards for policy gradient. Furthermore, we introduce the first image and poem pair dataset (MultiM-Poem) and a large poem corpus (UniM-Poem) to enhance researches on poem generation, especially from images. Extensive experiments demonstrated that our embedding model can approximately learn a rational visual-poetic embedding space. Objective and subjective evaluation results demonstrated the effectiveness of our poem generation model.

## ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number 18H03244 and 18H03243. The authors would like to thank Dr. Tao Mei for his insightful discussions on this work when he was a senior research manager at Microsoft Research, and thank Yuanchun Xu, Guang Zhou, Hongyuan Zhu, Yang Xiang and Yu Shi for promoting the technology transfer of this research to Microsoft Xiaoice.

## REFERENCES

- [1] Tseng-Hung Chen, Yuan-Hong Liao, Ching-Yao Chuang, Wan-Ting Hsu, Jianlong Fu, and Min Sun. 2017. Show, Adapt and Tell: Adversarial Training of Cross-domain Image Captioner. *ICCV* (2017), 521–530.
- [2] Xinlei Chen and C Lawrence Zitnick. 2015. Mind’s eye: A recurrent visual representation for image caption generation. In *CVPR*. 2422–2431.
- [3] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *NIPS 2014 Workshop on Deep Learning* (2014).
- [4] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. 2015. From captions to visual concepts and back. In *CVPR*. 1473–1482.
- [5] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *ECCV*. 15–29.
- [6] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. 2013. Devise: A deep visual-semantic embedding model. In *NIPS*. 2121–2129.
- [7] Jianlong Fu, Heliang Zheng, and Tao Mei. 2017. Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-grained Image Recognition. In *CVPR*. 4438–4446.
- [8] Marjan Ghazvininejad, Xing Shi, Yejin Choi, and Kevin Knight. 2016. Generating Topical Poetry.. In *EMNLP*. 1183–1191.
- [9] Marjan Ghazvininejad, Xing Shi, Jay Priyadarshi, and Kevin Knight. 2017. Hafez: an Interactive Poetry Generation System. *ACL* (2017), 43–48.
- [10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NIPS*. 2672–2680.
- [11] Jing He, Ming Zhou, and Long Jiang. 2012. Generating Chinese Classical Poems with Statistical Machine Translation Models.. In *AAAI*.
- [12] Jack Hopkins and Douwe Kiela. 2017. Automatically Generating Rhythmic Verse with Neural Networks. In *ACL*, Vol. 1. 168–178.
- [13] Long Jiang and Ming Zhou. 2008. Generating Chinese couplets using a statistical MT approach. In *COLING*. 377–384.
- [14] Andrej Karpathy, Armand Joulin, and Fei Fei Li. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *NIPS*. 1889–1897.
- [15] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539* (2014).
- [16] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *NIPS*. 3294–3302.
- [17] Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2017. A hierarchical approach for generating descriptive image paragraphs. *CVPR* (2017), 3337–3345.
- [18] 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 *CVPR*. 1601–1608.
- [19] 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 *AAAI*.
- [20] Hisar Maruli Manurung. 1999. A chart generator for rhythm patterned text. In *Proceedings of the First International Workshop on Literature in Cognition and Computer*. 15–19.
- [21] Hugo Oliveira. 2009. Automatic generation of poetry: an overview. *Universidade de Coimbra* (2009).
- [22] Hugo Gonçalo Oliveira. 2012. PoeTryMe: a versatile platform for poetry generation. *Computational Creativity, Concept Invention, and General Intelligence* 1 (2012), 21.
- [23] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*. 311–318.
- [24] Cesc C Park and Gunhee Kim. 2015. Expressing an image stream with a sequence of natural sentences. In *NIPS*. 73–81.
- [25] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *CVPR*, Vol. 1. 3.
- [26] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *IJCV* 115, 3 (2015), 211–252.
- [27] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
- [28] Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *HLT-NAACL*. 173–180.
- [29] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *CVPR*. 3156–3164.
- [30] Jingwen Wang, Jianlong Fu, Yong Xu, and Tao Mei. 2016. Beyond Object Recognition: Visual Sentiment Analysis with Deep Coupled Adjective and Noun Neural Networks.. In *IJCAI*. 3484–3490.
- [31] Limin Wang, Sheng Guo, Weilin Huang, and Yu Qiao. 2015. Places205-vggnet models for scene recognition. *arXiv preprint arXiv:1508.01667* (2015).
- [32] Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8, 3-4 (1992), 229–256.
- [33] 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 *AAAI*.
- [34] Zhen Xu, Bingquan Liu, Baoxun Wang, SUN Chengjie, Xiaolong Wang, Zhuoran Wang, and Chao Qi. 2017. Neural Response Generation via GAN with an Approximate Embedding Layer. In *EMNLP*. 628–637.
- [35] Rui Yan, Han Jiang, Mirella Lapata, Shou-De Lin, Xueqiang Lv, and Xiaoming Li. 2013. i, Poet: Automatic Chinese Poetry Composition through a Generative Summarization Framework under Constrained Optimization.. In *IJCAI*. 2197–2203.
- [36] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. 2017. Boosting image captioning with attributes. In *IEEE International Conference on Computer Vision, ICCV*. 22–29.
- [37] Xiaoyuan Yi, Ruoyu Li, and Maosong Sun. 2017. Generating Chinese Classical Poems with RNN Encoder-Decoder. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*. 211–223.
- [38] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *CVPR*. 4651–4659.
- [39] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient.. In *AAAI*. 2852–2858.
- [40] Wojciech Zaremba and Ilya Sutskever. 2015. Reinforcement Learning Neural Turing Machines-Revised. *arXiv preprint arXiv:1505.00521* (2015).
- [41] Xingxing Zhang and Mirella Lapata. 2014. Chinese Poetry Generation with Recurrent Neural Networks.. In *EMNLP*. 670–680.