

MSCap: Multi-Style Image Captioning with Unpaired Stylized Text

Longteng Guo^{1,4} Jing Liu*¹ Peng Yao² Jiangwei Li³ Hanqing Lu¹

¹National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences

²University of Science and Technology Beijing ³Multimedia Department, Huawei Devices

⁴University of Chinese Academy of Sciences

{longteng.guo, jliu, luhq}@nlpr.ia.ac.cn, S20180598@xs.ustb.edu.cn, lijiangwei1@huawei.com

Abstract

In this paper, we propose an adversarial learning network for the task of multi-style image captioning (MSCap) with a standard factual image caption dataset and a multi-stylized language corpus without paired images. How to learn a single model for multi-stylized image captioning with unpaired data is a challenging and necessary task, whereas rarely studied in previous works. The proposed framework mainly includes four contributive modules following a typical image encoder. First, a style dependent caption generator to output a sentence conditioned on an encoded image and a specified style. Second, a caption discriminator is presented to distinguish the input sentence to be real or not. The discriminator and the generator are trained in an adversarial manner to enable more natural and human-like captions. Third, a style classifier is employed to discriminate the specific style of the input sentence. Besides, a back-translation module is designed to enforce the generated stylized captions are visually grounded, with the intuition of the cycle consistency for factual caption and stylized caption. We enable an end-to-end optimization of the whole model with differentiable softmax approximation. At last, we conduct comprehensive experiments using a combined dataset containing four caption styles to demonstrate the outstanding performance of our proposed method.

1. Introduction

Automatically generating human-like captions for images, namely image captioning, has emerged as a prominent interdisciplinary research problem at the intersection of computer vision and natural language processing [36, 33, 40]. It has many important industrial applications,

*Corresponding Author

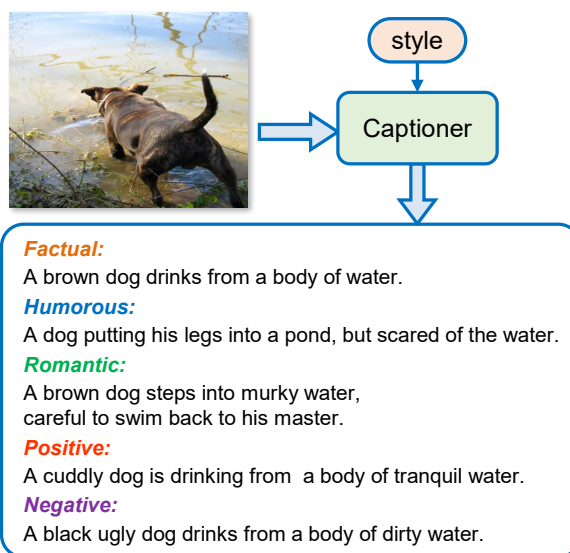


Figure 1. Example results of our multi-style image-captioning model. Given an image, our model learns to generate attractive image captions with various styles, which could be controlled by assigning style labels. Sentences under each colored words, i.e. the style name, are the generated caption corresponding to that style.

such as visual intelligence in chatting robots, photo sharing on social media, and assistive facilities for visually impaired people. To generate true human-like image captions, an image captioning system is required to understand the visual content and write captions with proper linguistic properties. However, most existing image captioning systems focus on the vision side that describes the visual content in an objective, neutral manner (*factual* captions), while the language side, e.g. linguistic style, is often neglected.

In fact, linguistic style [4] is an essential factor in human language that reflects personality, emotion, and sentiment. Style typically refers to linguistic aspects other than the

message content. Figure 1 show the captions of distinctive styles for a given image, including factual, humorous, romantic, positive, and negative. Incorporating appropriate styles into image captions will greatly enrich their clarity and attractiveness, and thus foster user engagement and social interactions. Some efforts have been made on stylized image captioning, including explicitly modeling sentiment words [25], transforming word embeddings matrices [10], and factoring the problem into two separate subprocesses [24] *et al.* However, all these models are built to translate images into captions of a single caption style. So far there has not been an efficient way to simultaneously handle multiple styles. Their inefficiency results from the fact that in order to learn mappings between images and k caption styles, k distinctive models have to be trained. Meanwhile, the model can only learn from a specific style out of k and cannot fully utilize the entire training data, even though there exists common knowledge that could be learned from the whole k -style data, *e.g.* correspondence between words and image content.

To address this problem, a single-model solution for multi-style image captioning (MSCap) is desired to generate visually grounded and any desired stylized captions for a given image, while multi-style captioning resources including images and multi-style captions are explored jointly for the single-model training. Typically, training such a model requires fully annotated collections of aligned image-stylized-caption pairs (paired data) for each style. However, it is quite expensive to collect such paired multi-style captioning collections, especially when the numbers of images and styles increase. Compared to annotating stylized captions for each image, it is much easier and cheaper to collect a corpus of stylized sentences without aligned images. Therefore, it is challenging but valuable to design a multi-style captioning model by exploring such unpaired multi-stylized data in addition to handily available factual image-caption paired data (*e.g.* MS COCO [22] dataset), which motivates our work.

In this paper, we propose an adversarial learning network to handle the problem of multi-style image caption generation simultaneously with factual image-caption pairs, and unpaired stylized captions. Given an image and its desired captioning style as input, the proposed model generates its corresponding stylized caption. Specifically, the proposed adversarial learning framework consists of five modules in which the first module is a typical image encoder, and the following four modules are the main focuses of this paper. First, we design a style dependent caption generator to output a sentence conditioned on an encoded image and a specified style. Second, a caption discriminator is presented to distinguish the input sentence is real or not. The discriminator is performed in an adversarial manner with the generator during training, and thus guides the generator towards

generating more natural and human-like caption. Third, a style classifier is introduced to discriminate what the specific style of the input sentence is. We further introduce a back-translation module to ensure that the generated stylized captions are visually grounded. The basic intuition is that there exists content consistency between a stylized caption and a factual caption describing the same image. Given a pair of image and factual caption, if we generate, *e.g.*, a humorous caption from the image, and then translate it into a factual caption, we should arrive at the real factual caption. We name this process back-translation and implement it via a multilingual neural machine translation (NMT) [14] model in which the multi-stylized captions are regarded as source languages, and the factual caption as the target language. Overview of the framework is illustrated in Figure 2. We enable an end-to-end optimization of the whole model with differentiable softmax approximation [13] which anneals smoothly to discrete case. At last, we conduct comprehensive experiments using a combined dataset containing five caption styles: humorous, romantic, positive, negative and factual styles. As far as our knowledge goes, our work is the first to successfully perform multi-style image captioning with unpaired stylized data. In summary, the main contributions of this paper are:

- We propose MSCap, a unified multi-style image captioning model that learns to map images into attractive captions of multiple styles. The model is end-to-end trainable without using supervised style-specific image-caption paired data.
- We design a novel style-dependent caption generator that which enables leveraging unpaired stylized captions for model pre-training. And we introduce a back-translation module to assure the generated captions to be consistent with the image content.
- We provide both qualitative and quantitative results on the multi-style and single-style image captioning tasks, showing the superiority of our proposed model.

2. Related Work

2.1. Image Captioning

Recent advances in deep learning and release of large scale datasets, *e.g.* MS COCO [22] and Flickr30k [27], have led to end-to-end trainable image captioning models. Most modern image captioning systems adopts the encoder-decoder framework [36, 40, 38, 41], where a convolutional neural network (CNN) encodes images into visual features, and a RNN takes the image features as inputs to decode them into sentences, typically trained end-to-end by maximum likelihood estimation. It has been shown that attention mechanisms [40, 23, 1] and high-level attributes/concepts

[42, 48] can help image captioning. Recently, reinforcement learning is introduced into image captioning models to directly optimize task-specific metrics [28, 46]. Some works adopt GANs to generate human-like [29] or diverse captions [21].

2.2. Stylized Image Captioning

Stylized image captioning aims at generating captions that are successfully stylized and describe the image content accurately. Some works have been proposed to tackle this task, which could be divided into two categories: models using parallel stylized image-caption data (supervised mode) [25, 7, 31, 43] and models using non-parallel stylized corpus (semi-supervised mode) [10, 24]. SentiCap [25] handles the positive/negative styles and proposes to model word changes with two parallel Long Short Term Memory networks (LSTM) and word-level supervisions. StyleNet [10] handles the humorous/romantic styles by factoring the input weight matrices to contain a style specific factor matrix. SF-LSTM [7] experiments on the above four caption styles and propose to learn two groups of matrices to capture the factual and stylized knowledge, respectively.

However, all these works are built to translate images into captions of a single caption style, while our model can simultaneously handle multiple styles. More similar to our work, You *et al.* [43] propose two simple methods to inject sentiments into image captions and can control the sentiment by providing different sentiment labels. However, this model is trained in supervised mode, while our model works in a harder semi-supervised mode with no requirement on parallel stylized data.

2.3. Generative Adversarial Networks

The Generative Adversarial Networks (GANs) [11] framework learn generative models without explicitly defining a loss function for a target distribution. GANs has shown promising results in fields of computer vision, including image super-resolution [20], photo editing [6, 30], domain adaptation [35, 5], image-to-image translation [26, 15, 9] and text-to-image translation [45]. Though GANs have achieved great successes on computer vision applications, there are only little progress on applying it to sequence generation tasks because the non-differentiability of discrete word tokens makes generator optimization difficult. Recently, some techniques have been proposed to address the non-differentiable challenge [19, 44, 13]. In our work, we employ the method proposed in [13] which uses continuous relaxation to approximate the discrete sampling process so that the training procedure can be effectively optimized through back-propagation.

3. MSCap for Multi-Style Image Captioning

We first present the overview of our MSCap framework (Sec. 3.1), then describe each module of it and introduce the objectives and strategy for training.

3.1. Framework Overview

The overall framework of the proposed MSCap is illustrated in Figure 2. It is comprised of five basic subnetworks, i.e., an image encoder E , a caption generator G , a caption discriminator D , a style classifier C , and a back-translation network T . We are given a factual dataset $\mathcal{P} = \{(x, \hat{y}_f)\}$, with paired image x along with its corresponding factual caption \hat{y}_f , and a collection of unpaired stylized sentences $\mathcal{P}^u = \{(\hat{y}_s, s)\}$, $s \in \{s_1 \dots s_k\}$ containing captions of k distinctive styles, where \hat{y}_s denote a stylized caption with style s . We regard the factual captions \hat{y}_f as having the ‘‘factual’’ style, denoted as s_0 , which would help model training since the large dataset of factual captions can be included in the training data. We denote the extended stylized corpus dataset as $\mathcal{P}' = \{(\hat{y}_s, s)\}$, $s \in \{s_0, \dots, s_k\}$. Given an image x and a style label s , we aim at generating a sentence y such that: 1) y is a natural sentence, 2) y is of style s , and 3) (x, y) forms a relevant pair.

The caption generator G conditions on the encoded image features $E(x)$ and a target style label s to generate a sentence y , i.e. $y = G(E(x), s)$. This sentence is fed into D , C , and T for enforcing it to satisfy the three requirements, respectively. Specifically, the discriminator D classifies whether a caption is a natural, human-like caption by distinguishing the fake generated caption y from real human-written captions $(\hat{y}_s, s \in \{s_0, \dots, s_k\})$. The style classifier C produces probability distributions of y belonging to each of the $k + 1$ style categories, and a style classification loss is thus calculated for enforcing y to be in the given style s . The back-translation module T ensures y is visually grounded on x . That is achieved by ‘‘translating’’ y back into \hat{y}_f (i.e. $T(y, s) \rightarrow \hat{y}_f$) in the sense of cycle-consistency [49]. The whole system is end-to-end trained by using differentiable softmax approximation in the caption generator.

3.2. Image Encoder

Given an image x , we first encode it to obtain image features using a deep CNN. The image features could be a static, global pooled representation of the image [37], or the spatial visual features [40]. Based on the features, a visual context vector is obtained for each time step, by directly using the static feature or calculating adaptively with the soft-attention mechanism [40] from the visual features. In this paper, we use the static feature to be consistent with precious works, thus the context vector c^v is $c^v = E(x)$.

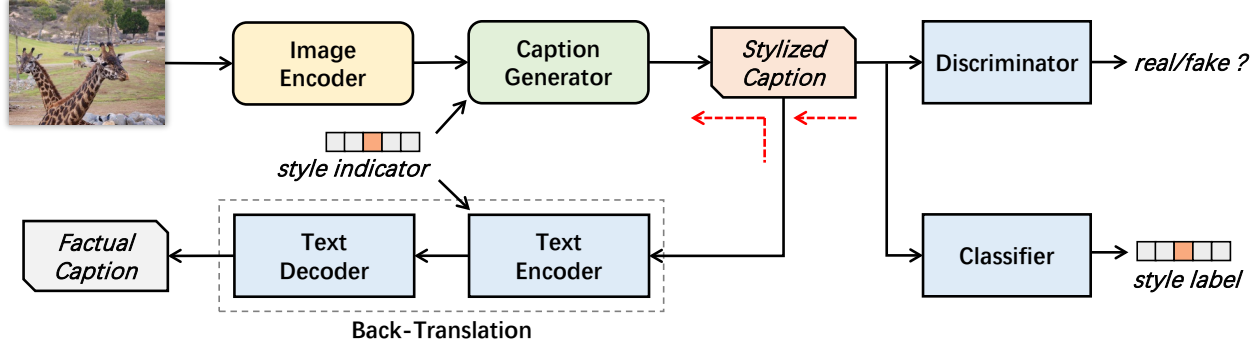


Figure 2. Overall framework of our MSCap. The multi-style caption generator takes in the encoded image feature and a style indicator as input to generate a caption with target style. The adversarial loss, style classification loss, and back-translation loss are then calculated based on the discriminator, classifier, and back-translation network, respectively. The red arrows denote gradient propagation enabled by the differentiable approximation.

3.3. Caption Generator

We design a style-dependent caption generator G that fully capture the language properties of each style by enabling directly training G with unpaired stylized captions.

Condition G on style labels. To effectively inject style conditions into G , we use an $(k + 1)$ -dimensional one-hot vector to represent the $k + 1$ different styles, with each element represents a corresponding style. we first feed s into a style embedding layer and then concatenate the resulting style embedding vector with the input word embedding vector as the input vector (w_t) to the LSTM at each step.

Enable training on unpaired corpus. For unpaired stylized corpus, its syntax and grammar rules are significantly different from that of the paired factual captions. Therefore, it’s beneficial to explicitly model the language properties of the unpaired corpus. However, current models usually adopt the “**injecting**” mode [34], which deeply couples the visual and linguistic information inside the recurrent loop of the RNN/LSTM, as is shown in Figure 3 (a). Such a mode fails to capture the language properties of unpaired corpus because the model cannot be trained without the presence of images.

To address this problem, we base G on the “**merging**” mode [34] and a **style gate** (as is shown in Figure 3 (b)). We first move the visual context out of the LSTM, leaving the LSTM modeling the linguistic information only. We then introduce an additional multimodal fusion module to merge the visual context c^v and linguistic context c_t^l for predicting words. The style gate provides the word predictor a fallback option to rely only on c_t^l when the image is unavailable. Inspired by [23], we design the style gate to adaptively assign

different weights to c^v and c_t^l :

$$g_t = \sigma(w_g^T \tanh(W_g[c_t^l; h_t] + b_g)), \quad (1)$$

$$c_t = g_t c_t^l + (1 - g_t) c^v, \quad (2)$$

where $[\cdot]$ indicates concatenation, c_t is the mixed context vector, h_t is hidden state of LSTM, and σ is the sigmoid activation. c_t^l is calculated by $l_t = \sigma(W_l[w_t; c^v; h_t] + b_l)$, $c_t^l = l_t \odot \tanh(m_t)$, where l_t is a gate vector, m_t is the memory cell state of the LSTM, w_t is the input vector, σ is the sigmoid activation and \odot represents element-wise product. A higher g_t means more focus on the linguistic context. Finally, the mixed context vector c_t is concatenated with the hidden state h_t and is then fed into the word classifier to produce the probability over the vocabulary of possible words:

$$p_t = \text{softmax}\left(\frac{W_o[c_t; h_t]}{\tau}\right), \quad (3)$$

where $\tau \in (0, 1)$ a temperature parameter. When training with the unpaired stylized corpus, it is natural to turn the style gate only to the linguistic context vector, *i.e.* $g_t = 1, c_t = c_t^l$. In this case, the model relies totally on the linguistic context for word prediction, and becomes a pure language model.

We pre-train the caption generator with both paired factual data \mathcal{P} and unpaired stylized corpus \mathcal{P}^u by maximizing the log-likelihood of the ground-truth captions:

$$\theta^* = \arg \max_{\theta} \mathbb{E}_{(x, \hat{y}_f) \in \mathcal{P}} \log p(\hat{y}_f | x, s_0; \theta) + \mathbb{E}_{(\hat{y}_s, s) \in \mathcal{P}^u} \log p(\hat{y}_s | s; \theta), \quad (4)$$

where θ is the parameters of G and s_0 denotes the factual style.

3.4. Adversarial Loss

To make the generated captions indistinguishable from real captions, we adopt adversarial training with a discrim-

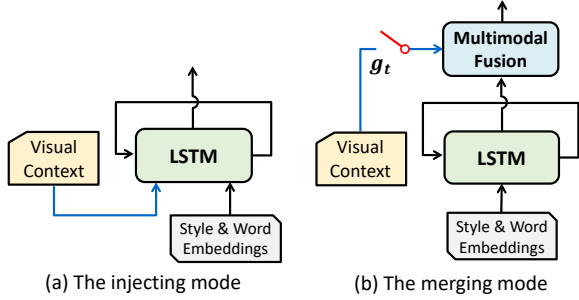


Figure 3. Comparison between the injecting and merging modes. Our generator cooperates the merging mode with the style gate (g_t in (b)), which enables training directly with unpaired corpus.

inator D , where G generates a fake caption $G(x, s)$ and D tries to distinguish it from real captions. The adversarial loss [11] is calculated by:

$$\mathcal{L}_{adv} = \mathbb{E}_{\hat{y}}[\log D(\hat{y})] + \mathbb{E}_{x,s}[1 - \log D(G(x, s))], \quad (5)$$

where \hat{y} is a real caption from \mathcal{P}' , x is an image from \mathcal{P} , and s is a style label randomly sampled from $\{s_0, \dots, s_k\}$. G tries to minimize this objective, while D tries to maximize it.

3.5. Style Classification Loss

Given an image x and a target style label s , it is required that the generated caption should correctly own the target style. To satisfy this condition, we employ an style classifier C to constrain the generated caption y to own the desired style, *i.e.* $C(G(x, s)) \rightarrow s$. The style classification loss for C and G is formulated as follows:

$$\mathcal{L}_{cls} = \mathbb{E}_{\hat{y}}[-\log C(s_0|\hat{y})] + \mathbb{E}_{x,s}[-\log C(s|G(x, s))]. \quad (6)$$

3.6. Back-Translation Loss

By minimizing the adversarial and classification losses (Eqn. 5 and 6), G is trained to generate captions that are human-like and classified to its correct target style. However, minimizing the two losses along does not guarantee that generated captions accurately describe the content of its input images, *i.e.* visually grounded. To alleviate this problem, we introduce the back-translation module T to impose a condition on the relation among y , \hat{y}_f , and x .

We begin from the observation that the factual image-caption pair (x, \hat{y}_f) shares the same content information. From this point, the relevancy between the generated caption y and the image x can be approximated by the relevancy between y and the ‘‘ground-truth’’ factual caption \hat{y}_f . Thus, we constrain y to be consistent with \hat{y}_f in the sense of sentence content. This is achieved by using the back-translation module T that ‘‘translates’’ y back into y_f , *i.e.*

$T(y, s) \rightarrow \hat{y}_f$. T is implemented as a multilingual neural machine translation (NMT) network in which the multi-stylized captions are regarded as source languages, and the factual caption as the target language. Concretely, T includes a text encoder that takes y and the target style labels s as inputs, and a followed text decoder that takes the outputs of the text encoder as input to generate a sentence. We then formulate the back-translation loss as minimizing the negative log-likelihood of the factual caption:

$$\mathcal{L}_{trans} = \mathbb{E}_{(x, \hat{y}_f), s}[-\log p(\hat{y}_f|G(x, s), s; T)]. \quad (7)$$

Another possible method to enforce cycle-consistency is directly translating y back into the image x (or image features $E(x)$) [45, 47]. However, the text-to-image synthesis itself is a tough task and so far the performance is far from satisfaction. While translation between two sentences is much more mature and practical.

3.7. Full Objectives

Finally, the objective functions for G , D , C , and T are written, respectively, as

$$\begin{aligned} \mathcal{L}_G &= -\lambda_{adv}\mathcal{L}_{adv} + \lambda_{cls}\mathcal{L}_{cls} + \lambda_{trans}\mathcal{L}_{trans}, \\ \mathcal{L}_D &= \mathcal{L}_{adv}, \quad \mathcal{L}_C = \mathcal{L}_{cls}, \quad \mathcal{L}_T = \mathcal{L}_{trans}, \end{aligned} \quad (8)$$

where λ_{adv} , λ_{cls} and λ_{trans} are hyper-parameters for balancing the losses.

3.8. Training Strategy

Adversarial training over the discrete samples generated by G hinders gradients propagation. Although sampling-based gradient estimator such as REINFORCE [39, 44] can be adopted, we found that training with these methods can be unstable due to the high variance of the gradient and also inefficient since Monte Carlo roll-out is often required. Instead, we employ the continuous approximation technique proposed by Hu *et al.* [13] to enable end-to-end optimization of the whole model.

Specifically, instead of sampling a single *hard* word (one-hot vector) from p_t (Eqn. 3), we consider the peaked distribution vector p_t itself as a *soft* word, which is the output of G at the t -th step and serves as an input in the $t+1$ -th step. At the $(t+1)$ -th step, we compute the word embedding vector with $e_{t+1} = W_e p_t$, where $p_t \in \mathbb{R}^N$, $e_{t+1} \in \mathbb{R}^d$, and $W_e \in \mathbb{R}^{d \times N}$ is the word embedding matrix. e_{t+1} is then fed into the LSTM. The temperature τ gradually anneals to 0 (the discrete case) as training proceeds. We empirically find that this simple yet effective approach enjoys low variance and fast convergence. In practice, we employ the Wasserstein GAN [2] for optimizing the adversarial loss \mathcal{L}_{adv} .

4. Experimental Setup

4.1. Dataset

We conduct experiments on two publicly available stylized image caption datasets, FlickrStyle10K [10] and SentiCap [25], and a large factual image-caption dataset, MS COCO [22]. **COCO** is a large image captioning dataset, containing 82783, 40504 and 40775 images for training, validation and test, respectively, with 5 factual captions for each image. **FlickrStyle10K** contains 10K Flickr images with stylized captions. However, only the 7K training set are public, in which each image is labeled with 5, 1, and 1 captions for factual, humorous, and romantic styles, respectively. Following [7], we randomly select 6,000 and 1,000 of them as the training and test sets, respectively. **SentiCap** is an image sentiment captioning dataset based on COCO images, which contains images that are labeled by 3 positive and 3 negative sentiment captions. The positive and negative subsets contain 998/673 and 997/503 images for training/testing, respectively. We randomly sample 100 images from each of the training splits for evaluation. For convenience, we denote the humorous, romantic, positive, negative styles, and factual as **Humor**, **Roman**, **Pos**, **Neg**, and **Fact**, respectively. For stylized data, during training, only the captions from the training split are used, while when testing, both the images and captions from the test split are used for benchmarking the models. The training set of COCO is used as the paired factual dataset \mathcal{P} while the captions from all the five styles are used as the unpaired stylized corpus \mathcal{P}' .

4.2. Compared Approaches

There are only few works that address the stylized image captioning problem with unpaired data (semi-supervised learning) as ours do. Thus, we also compare our model with models using paired training data, *i.e.* learning in fully-supervised mode. We compare our approach with the following methods:

- NIC [36]: the standard encoder-decoder model. We train it with factual image-caption pairs from COCO and treat it as the factual baseline.
- NIC-FT: We finetune the trained NIC model with paired stylized data on each of the four styles separately.
- SF-LSTM [7]: the current state-of-the-art supervised model for single-style image captioning.
- StyleNet [10]: the single-style semi-supervised model that factors the input weight matrices to contain a style specific factor matrix. We implement this model to first pre-train it with paired factual data and then separately train four models for each style.

4.3. Implementation Details

We extract the 2048-dimensional image features from the last pooling layer of ResNet-101 [12]. The dimensions of the caption generator’s LSTM hidden states and word embeddings are fixed to 512 for all of the models discussed herein. The dimensions of the style embeddings are set to 20. The discriminator D and classifier C are implemented as CNNs [16] with highway connections [17]. The back-translation network T is built on two Gated Recurrent Unit (GRU) [8] networks, which are used as the text encoder and decoder, respectively. The global attention mechanism [3] is adopted in the decoder to decide which part of the source sentence to pay attention to. All the sub-networks share the same word embedding and style embedding. We first pre-train the generator using both the paired factual image-caption data and unpaired stylized corpus (Eqn. 4), with an initial learning rate of 5×10^{-4} .

After that, we train the whole network, including G , D , C , and T , all together according to Eqn. 8. We use ADAM [18] optimizer for all the sub-networks, and use fixed learning rates of 5×10^{-5} for G , D , C and 5×10^{-4} for T . We train D for 5 times more than G . We use a mini-batch size of 80. Beam search with a beam size of 3 is used when testing. We use a fixed temperature τ of 0.1. We set λ_{adv} , λ_{cls} , and λ_{trans} to 0.2, 1, and 5, respectively.

5. Experimental Results

5.1. Quality of Generated Captions

We evaluate the quality of generated captions in terms of relevance with input images, fluency, and accuracy of style.

Relevancy For each of the five styles, the image-stylized caption pairs in the testing split could be used for benchmarking the models [25, 7]. We report the widely used automatic evaluation metrics, BLEU-1, BLEU-3, METEOR, and CIDEr [22]. These metrics are mostly based on n-gram overlap, which are not perfect metrics for evaluating stylized captions because stylized image captioning allows more flexibility for choosing words and phrases used to describe an image. Table 1 and 2 summarize the results on the Pos/Neg and Roman/Humor styles, respectively. Compared with the semi-supervised model, *i.e.* StyleNet, our multi-style model achieves the best performance on all styles, including Pos, Neg, Roman, and Humor. Compared with fully-supervised models, our model is close to these models on the Pos/Neg styles. While on the harder Roman/Humor styles, the scores are lower because the humor/roman captions are typically much longer and more flexible. Specifically, our model gets comparable scores on BLEU-1, while its BLEU-3 score is lower. That is corresponding to our intuition: because BLEU-n measures the precision and recall

Table 1. Performance comparisons on the test splits of Pos and Neg styles. unpaired means the model uses unpaired stylized text for training, *i.e.* semi-supervised learning. B@n, M, C, ppl., cls. are short for BLEU-n, METEOR, CIDEr, perplexity, style classification accuracy (%), respectively. For ppl. smaller is better, for the others larger is better.

Model	Un-paired	Multi-style	Positive						Negative					
			B@1	B@3	M	C	ppl.	cls.	B@1	B@3	M	C	ppl.	cls.
NIC	no	no	47.6	16.3	14.9	55.1	25.6	22.4	46.9	16.1	14.8	54.0	25.4	23.2
NIC-FT	no	no	48.2	17.3	16.6	54.3	20.4	91.3	47.3	17.8	16.1	55.4	21.5	89.5
SF-LSTM	no	no	50.5	19.1	16.6	60.0	–	–	50.3	20.1	16.2	59.7	–	–
StyleNet	yes	no	45.3	12.1	12.1	36.3	24.8	45.2	43.7	10.6	10.9	36.6	25.0	56.6
MSCap	yes	yes	46.9	16.2	16.8	55.3	19.6	92.5	45.5	15.4	16.2	51.6	19.2	93.4

Table 2. Performance comparisons on the test splits of Roman and Humor styles.

Model	Un-paired	Multi-style	Romantic						Humorous					
			B@1	B@3	M	C	ppl.	cls.	B@1	B@3	M	C	ppl.	cls.
NIC	no	no	25.1	7.0	10.6	33.0	61.6	24.3	25.5	7.2	9.7	33.5	57.1	25.5
NIC-FT	no	no	26.9	7.5	11.0	35.4	27.7	82.6	26.3	7.4	10.2	35.1	31.8	80.1
SF-LSTM	no	no	27.8	8.2	11.2	37.5	–	–	27.4	8.5	11.0	39.5	–	–
StyleNet	yes	no	13.3	1.5	4.5	7.2	52.9	37.8	13.4	0.9	4.3	11.3	48.1	41.9
MSCap	yes	yes	17.0	2.0	5.4	10.1	20.4	88.7	16.3	1.9	5.3	15.2	22.7	91.3

of n-grams, it is too hard for a semi-supervised model to achieve the exact matching of long phrases, *e.g.* 3-grams.

Fluency We evaluate the fluency of the generated captions in terms of the target style. We use a language modeling toolkit, SRILM [32], to test the fluency of generated sentences. SRILM calculates the perplexity of the generated sentences using the trigram language model trained on the respective corpus. We train such language models on each of the stylized corpus and compute the perplexity scores (denoted as **ppl.**) for the generated captions of each style and each model. Lower perplexity score of a caption indicates it is more fluent and appropriately stylized. The results are shown in Table 1 and 2 (see ppl. columns). As we can see, our approach maintains the lowest perplexity scores across all styles, including the supervised models. Particularly, our model maintains significantly better fluency than StyleNet.

Style accuracy We measure how often a generated caption has the correct target style according to a pre-trained style classifier. For this purpose, we use the TextCNN [16] as a style judge. It’s trained on the \mathcal{P}' dataset and achieves nearly perfect accuracy of 97.8%. The results of the style classification accuracy (denoted as **cls.**) are shown in Table 1 and 2 (see cls. columns). As can be seen, across all styles, our model achieves the highest style classification accuracy among all methods, including the oracle method, NIC-FT.

Table 3. Ablation study results. The scores of each metrics are its average scores on four styles (*e.g.* Pos, Neg, Humor, Roman).

Model	Cider	Perplexity ↓	Style acc.%
NIC	43.9	42.4	23.9
NIC-FT	45.1	25.4	85.9
StyleNet	22.9	37.7	45.4
MSCap	33.1	20.5	91.5
MSCap w/o adv.	14.6	57.1	66.7
MSCap w/o cls.	20.5	49.6	30.0
MSCap w/o trans.	7.72	13.6	96.0
MSCap w/o XE.	30.3	22.2	88.7

Human evaluations Automatic evaluation metrics cannot perfectly reflect the stylized captions’ quality in the users’ minds. Therefore, we perform human evaluation on the generated captions in terms of fluency, relevancy and style appropriateness. We randomly selected 50 images from the testing set and generate stylized captions for each image, resulting totally 50×4 image-caption pairs to be evaluated. We asked 10 volunteers to rate the captions. The volunteers were asked to rank the generated captions in terms of their fluency, relevancy, and style appropriateness. Fluency was rated from 0 (unreadable) to 3 (perfect). Relevancy was rated from 0 (unrelated) to 3 (very related). Style appropriateness means whether a caption appropriately owns the desired styles, rated from 0 (bad) to 3 (perfect). The scores on each style and their average are shown in Table 4. As we can see, our MSCap rates between 1.92 ~ 2.62




				
Factual	a man smiles near people as he skateboards indoors.	two giraffes are standing outdoors near a building.	a man riding skis down a snow covered slope.	an elephant is in some brown grass and some trees.
Roman	a man jumping a skateboard in the room, proud of his accomplishment.	two giraffes are walking through the filed, exploring the woods.	a man in a black jacket is jumping over a snow covered mountain to experience the thrill of life.	a baby elephant is running through the grass to meet his lover.
Humor	a man does tricks on his skateboard to show off.	two hungry giraffes standing in the filed looking for things to eat.	a lonely skier goes down an snowy hill thinking of cute lady skiers.	a elephant is balancing on a grass covered field.
Pos	a great image of a young people do tricks on his skateboards.	two giraffes in a pleasant park are against beautiful trees.	a amazing people stand on his skis on a snowy hill.	an elephant enjoying the nice day while standing on the grass.
Neg	a poor boy skateboards in the crowded room.	two giraffes are against a broken tree and dead grass.	a man stands on his broken skis in the dirty snow.	a poor elephant is approaching a dead filed.

Figure 4. Examples of the stylized captions generated by MSCap. Each column shows an image and its corresponding captions, while captions at each row correspond to one of the caption styles: factual, romantic, humorous, positive, and negative.

Table 4. Human evaluations results of the generated captions in terms of fluency, relevancy, and style appropriateness.

Style	Pos	Neg	Roman	Humor	Avg.
Fluency	2.62	2.43	2.12	2.04	2.30
Relevancy	2.46	2.37	2.02	1.92	2.19
Style	2.33	2.28	2.12	2.06	2.20

on all the items among all styles, which could be considered satisfactory since the highest score is 3.

5.2. Ablation Study

We conduct ablation study to show how much each component of MSCap contributes to the caption quality. Specifically, we remove the adversarial loss (\mathcal{L}_{adv}), the classification loss (\mathcal{L}_{cls}) and the back-translation loss (\mathcal{L}_{trans}) from the generator’s objective function (\mathcal{L}_G in Eqn. 8), denoted as *w/o adv.*, *w/o cls.*, and *w/o trans.*, respectively. To show the effect of our designed generator that enables training directly with unpaired text (Eqn. 4), we train another MSCap model that only use paired data during the XE training, denoted as *w/o XE.*. The results are summarized in Table 3. As we can see, without \mathcal{L}_{adv} , the model performs very poorly in almost all metrics. We found that its output sentences are mostly non-fluent, which contains many repetitive words, such as “nice nice day”, “a a boy”. Without \mathcal{L}_{cls} , the model scores very low on the style classification accuracy, indicating that it fails to generate stylized captions of desired style. Without \mathcal{L}_{trans} , though the model scores the lowest perplexity and highest style accuracy scores, however, the CIDEr score decreases markedly. This is because although the individual sentences are fluent and

stylized, however, the captions are not event related to the images. Also, we found a large number of captions are identical. The results validate the significance and effectiveness of the back-translation module to enforce relevancy between generated captions and images. Without pre-training on unpaired stylized text (*w/o XE.*), the performance of the model drops on all metrics. We infer that pre-training on unpaired stylized text helps the generator to better capture the language properties of the stylized data.

5.3. Example Results

In Figure 4, we show four example captions generated by our MSCap. We can see that the captions are fluent, relevant to the image, and also correctly stylized with the target style. For example, the captions of the first image contain words (“proud”, “trick”, “great”, and “poor”) that match well with the desired styles (factual, romantic, humorous, positive, and negative styles, respectively).

6. Conclusion

We have proposed the MSCap, a multi-style image captioning model trained using unpaired stylized corpus. MSCap can generate human-like, appropriately stylized, visually grounded, and style-controllable captions. In addition, MSCap is a single unified model that can be easily scaled to more caption styles. Extensive experiments demonstrated the efficacy of MSCap.

Acknowledgment

This work was supported by National Natural Science Foundation of China (61872366 and 61472422) and Beijing Natural Science Foundation (4192059).

References

- [1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and vqa. *arXiv preprint arXiv:1707.07998*, 2017.
- [2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.
- [3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
- [4] Allan Bell. Language style as audience design. *Language in society*, 13(2):145–204, 1984.
- [5] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 7, 2017.
- [6] Andrew Brock, Theodore Lim, James M Ritchie, and Nick Weston. Neural photo editing with introspective adversarial networks. *arXiv preprint arXiv:1609.07093*, 2016.
- [7] Tianlang Chen, Zhongping Zhang, Quanzeng You, Chen Fang, Zhaowen Wang, Hailin Jin, and Jiebo Luo. ”factual” or ”emotional”: Stylized image captioning with adaptive learning and attention. *arXiv preprint arXiv:1807.03871*, 2018.
- [8] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.
- [9] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *arXiv preprint*, 1711, 2017.
- [10] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. Stylenet: Generating attractive visual captions with styles. In *Proc IEEE Conf on Computer Vision and Pattern Recognition*, pages 3137–3146, 2017.
- [11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *computer vision and pattern recognition*, pages 770–778, 2016.
- [13] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. Toward controlled generation of text. *arXiv preprint arXiv:1703.00955*, 2017.
- [14] Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. Google’s multilingual neural machine translation system: enabling zero-shot translation. *arXiv preprint arXiv:1611.04558*, 2016.
- [15] Taeksoo Kim, Moonsoo Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192*, 2017.
- [16] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.
- [17] Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. Character-aware neural language models. In *AAAI*, pages 2741–2749, 2016.
- [18] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [19] Matt J Kusner, Brooks Paige, and José Miguel Hernández-Lobato. Grammar variational autoencoder. *arXiv preprint arXiv:1703.01925*, 2017.
- [20] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, volume 2, page 4, 2017.
- [21] Dianqi Li, Xiaodong He, Qiuyuan Huang, Ming-Ting Sun, and Lei Zhang. Generating diverse and accurate visual captions by comparative adversarial learning. *arXiv preprint arXiv:1804.00861*, 2018.
- [22] Tsungyi Lin, Michael Maire, Serge J Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence Zitnick. Microsoft coco: Common objects in context. *european conference on computer vision*, pages 740–755, 2014.
- [23] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 6, 2017.
- [24] Alexander Mathews, Lexing Xie, and Xuming He. Semstyle: Learning to generate stylised image captions using unaligned text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8591–8600, 2018.
- [25] Alexander Patrick Mathews, Lexing Xie, and Xuming He. Senticap: Generating image descriptions with sentiments. In *AAAI*, pages 3574–3580, 2016.
- [26] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. *arXiv preprint arXiv:1610.09585*, 2016.
- [27] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.
- [28] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. *computer vision and pattern recognition*, 2017.
- [29] Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. Speaking the same language: Matching machine to human captions by adversarial training. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

- [30] Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, and Dimitris Samaras. Neural face editing with intrinsic image disentangling. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 5444–5453. IEEE, 2017.
- [31] Kurt Shuster, Samuel Humeau, Hexiang Hu, Antoine Bordes, and Jason Weston. Engaging image captioning via personality. *arXiv preprint arXiv:1810.10665*, 2018.
- [32] Andreas Stolcke. Srilm—an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*, 2002.
- [33] Jinhui Tang, Xiangbo Shu, Zechao Li, Guo-Jun Qi, and Jingdong Wang. Generalized deep transfer networks for knowledge propagation in heterogeneous domains. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 12(4s):68, 2016.
- [34] Marc Tanti, Albert Gatt, and Kenneth P Camilleri. Where to put the image in an image caption generator. *Natural Language Engineering*, 24(3):467–489, 2018.
- [35] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 4, 2017.
- [36] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3156–3164. IEEE, 2015.
- [37] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):652–663, 2017.
- [38] Jing Wang, Jianlong Fu, Jinhui Tang, Zechao Li, and Tao Mei. Show, reward and tell: Automatic generation of narrative paragraph from photo stream by adversarial training. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [39] Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3-4):229–256, 1992.
- [40] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron C Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *international conference on machine learning*, pages 2048–2057, 2015.
- [41] Zhilin Yang, Ye Yuan, Yuexin Wu, Ruslan Salakhutdinov, and William W Cohen. Review networks for caption generation. 2016.
- [42] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. 2016.
- [43] Quanzeng You, Hailin Jin, and Jiebo Luo. Image captioning at will: A versatile scheme for effectively injecting sentiments into image descriptions. *arXiv preprint arXiv:1801.10121*, 2018.
- [44] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In *AAAI*, pages 2852–2858, 2017.
- [45] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *arXiv preprint*, 2017.
- [46] Li Zhang, Flood Sung, Feng Liu, Tao Xiang, Shaogang Gong, Yongxin Yang, and Timothy M Hospedales. Actor-critic sequence training for image captioning. *arXiv preprint arXiv:1706.09601*, 2017.
- [47] Wei Zhao, Wei Xu, Min Yang, Jianbo Ye, Zhou Zhao, Yabing Feng, and Yu Qiao. Dual learning for cross-domain image captioning. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 29–38. ACM, 2017.
- [48] Luowei Zhou, Chenliang Xu, Parker Koch, and Jason J Corso. Image caption generation with text-conditional semantic attention. *arXiv preprint arXiv:1606.04621*, 2016.
- [49] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint*, 2017.