

Similar Scenes arouse Similar Emotions: Parallel Data Augmentation for Stylized Image Captioning

Guodun Li, Yuchen Zhai, Zehao Lin, Yin Zhang*

College of Computer Science and Technology, Zhejiang University, Hangzhou, China

{guodun.li,zhaiyuchen,georgelin,zhangyin98}@zju.edu.cn

ABSTRACT

Stylized image captioning systems aim to generate a caption not only semantically related to a given image but also consistent with a given style description. One of the biggest challenges with this task is the lack of sufficient paired stylized data. Many studies focus on unsupervised approaches, without considering from the perspective of data augmentation. We begin with the observation that people may recall similar emotions when they are in similar scenes, and often express similar emotions with similar style phrases, which underpins our data augmentation idea. In this paper, we propose a novel Extract-Retrieve-Generate data augmentation framework to extract style phrases from small-scale stylized sentences and graft them to large-scale factual captions. First, we design the emotional signal extractor to extract style phrases from small-scale stylized sentences. Second, we construct the pluggable multi-modal scene retriever to retrieve scenes represented with pairs of an image and its stylized caption, which are similar to the query image or caption in the large-scale factual data. In the end, based on the style phrases of similar scenes and the factual description of the current scene, we build the emotion-aware caption generator to generate fluent and diversified stylized captions for the current scene. Extensive experimental results show that our framework can alleviate the data scarcity problem effectively. It also significantly boosts the performance of several existing image captioning models in both supervised and unsupervised settings, which outperforms the state-of-the-art stylized image captioning methods in terms of both sentence relevance and stylishness by a substantial margin.

CCS CONCEPTS

• Information systems → Multimedia content creation.

KEYWORDS

stylized image captioning, data augmentation

ACM Reference Format:

Guodun Li, Yuchen Zhai, Zehao Lin, Yin Zhang. 2021. Similar Scenes arouse Similar Emotions: Parallel Data Augmentation for Stylized Image Captioning. In *Proceedings of the 29th ACM International Conference on Multimedia*

*Corresponding author: Yin Zhang

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 ACM 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.

MM '21, October 20–24, 2021, Virtual Event, China

© 2021 Association for Computing Machinery.

ACM ISBN 978-1-4503-8651-7/21/10...\$15.00

<https://doi.org/10.1145/3474085.3475662>

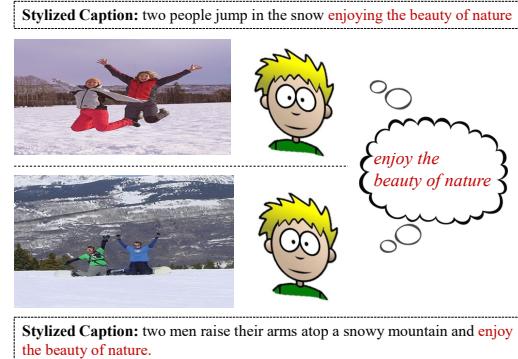


Figure 1: An example that when people are facing similar scenes, they arouse similar emotions, and express similar emotions with similar style phrases. A group of words in red is the style phrase of the stylized caption.

(MM '21), October 20–24, 2021, Virtual Event, China. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3474085.3475662>

1 INTRODUCTION

Stylized image captioning [8, 24] aims to generate a natural language description that is semantically related to a given image and consistent with a given style description. Since most image captioning models are data-hungry, one of the biggest challenges with this task is the lack of a large-scale parallel dataset containing dozens of pairs of an image and its stylized caption. It is time consuming and expensive to manually compose stylized captions for large-scale images. Hence, previous studies focus on unsupervised approaches, relying on unpaired stylized texts and existing large-scale factual image-caption dataset.

Limitations of such unsupervised approaches are as follows: 1) Most of the existing works [4, 23] learn the latent style vector based on unpaired stylized texts. However, in real-world applications, the style domain of large available unpaired texts are often far away from target style domain, making the trained models not able to perform well. 2) These well-designed unsupervised approaches are hard to employ state-of-the-art multi-modal pre-trained models [20, 45], which seriously limits the performance of this task. 3) It is hard to balance the use of the factual information in the large-scale image-caption corpus and style knowledge acquired from small-scale style texts because of the imbalanced data.

Contrasted with previous methods, we investigate how to improve the performance of this task from a different and orthogonal angle, i.e. augmenting more high-quality paired stylized data to alleviate the data scarcity problem. Our work begins with the real

experience during social interactions, where people arouse similar emotions when they are facing similar scenes, which is called the “context effect” [1, 7, 9] in psychology. This is because a similar visual context makes people extract similar emotions from past memory [9]. Further, people often express similar emotions with similar stylized phrases. Figure 1 shows two image-caption pairs, those images are similar to each other. And both captions share the same style phrase “*enjoy the beauty of nature*” to express the same emotion.

Based on the “context effect”, we propose a novel *Extract-Retrieve-Generate* data augmentation framework (ERG) to yield more paired stylized data in an automated manner. ERG is composed of three modules: 1) *Style phrase extractor* discerns linguistic style phrases from existing stylized captions in the small parallel stylized caption dataset. For example, for an input stylized caption “*two men raise their arms atop a snowy mountain and enjoy the beauty of nature*.”, the module is able to extract the style phrase “*enjoy the beauty of nature*”. 2) *Multi-modal scene retriever* returns a list of similar scenes in the small stylized data by leveraging the state-of-the-art multi-modal transformer-based models to compute the image-to-image, image-to-text, text-to-image, text-to-text similarities between sample pairs in the large factual corpus and those in the small stylized corpus, inspired by the observation that people often use similar images, texts, and other perceptual inputs as clues to recall past similar scenes. 3) *Emotion-aware caption generator* produces stylized image captions, given images in the large-scale factual dataset and the style phrases of scenes retrieved in the small stylized corpus.

The main contributions of this work are summarized as follows:

- To the best of our knowledge, this work is the first effort towards data augmentation of stylized image captioning. Our approach can be combined with any existing supervised, unsupervised and transformer-based approaches.
- We propose a novel Extract-Retrieve-Generate data augmentation framework to automatically generate high-quality and diversified stylized captions for images in the large-scale factual dataset by leveraging the visually or textually relevant scenes and their style phrases in the stylized corpus.
- Comprehensive experimental results on two widely-used stylized image-caption datasets show that our framework can alleviate the data scarcity problem significantly. It also effectively boosts the performance of several existing image captioning models in both supervised and unsupervised settings, which outperforms the state-of-the-art stylized image caption methods by a substantial margin.

2 RELATED WORK

2.1 Stylized Image Captioning

Image captioning is a hot and fundamental topic [10, 14, 30, 32] of intersection of advances in computer vision and natural language processing. Recently, novel image caption models [36, 41, 42] can generate captions from both visual latent space or multi-modal (images and texts) latent space by analyzing the visual content of the image first, e.g. using pre-trained image model like ResNet [34], and then generate image captions from the visual content using a language model [39, 40, 44].

Stylized image captioning [24] takes a further step, aiming at generating captions in target styles. Most existing work [4, 8, 11, 23, 43, 47] use non-parallel stylized image-caption data and the others [5, 24, 31] use parallel stylized corpus. StyleNet [8] proposes to generate humorous and romantic captions by factorizing the input weight matrices to contain a style-specific factor matrix. This work mainly focuses on capturing multiple linguistic styles information. Similarly, SF-LSTM [5] proposes to learn two groups of matrices to capture the factual and stylized knowledge, respectively. You et al.[43] choose another way to inject sentiments into image captions and control the sentiment by providing different sentiment labels in semi-supervised mode without requirement of parallel stylized data. SemStyle [23] develops a model that learns to generate visually relevant styled captions from a large corpus of stylized text without aligned images. Nezami et al.[25] propose an image captioning model called ATTEND-GAN which has an attention-based caption generator and adversarial training mechanism to generate correlated captions as well as more human-like variability of stylistic patterns. Similarly, Guo et al. [11] propose an adversarial learning framework with several auxiliary modules on unpaired stylized texts to generate captions in multiple styles. Recent advances in this task is MemCap[47], which explicitly encodes the knowledge about linguistic styles learned from unpaired stylized corpus with memory mechanism. Contrasted with previous works, we investigate how to improve the existing models from a different angle, i.e. producing more parallel stylized data with data augmentation techniques.

2.2 Data Augmentation

Data Augmentation is an effective way to improve the performance of neural models, which encompasses a set of techniques to increase the size and improve the quality of datasets. Such techniques reduce the cost of dataset collection, and usually boost the model performance on desired tasks. It has been widely applied in computer vision, natural language processing, and multimedia areas, especially for tasks without sufficient data [18, 22, 46]. In image captioning, data augmentation has been used in the previous work [16, 38], which focus on the image transformations to avoid over-fitting.

However, for stylized image captioning, due to the lack of parallel data, many studies focus on unsupervised approaches, which seriously limits the performance of powerful models such as pre-trained transformers [12, 20, 45]. There is no related work concerning data augmentation. Existing oversimple augmentation techniques based on a single modal such as image transformations can not contribute to the linguistic style of captions. To solve this pain point, we propose the Extract-Retrieve-Generate framework for paired stylized data augmentation in an automatic manner.

3 APPROACH

3.1 Overview

A stylized image captioning model is expected to generate a caption not only semantically related to the given image but also consistent with a given linguistic style. Suppose we have an insufficient training set $\mathcal{D}_s = \{(x_i^s, y_i^s)|_i\}$ of paired stylized data, where x_i^s and

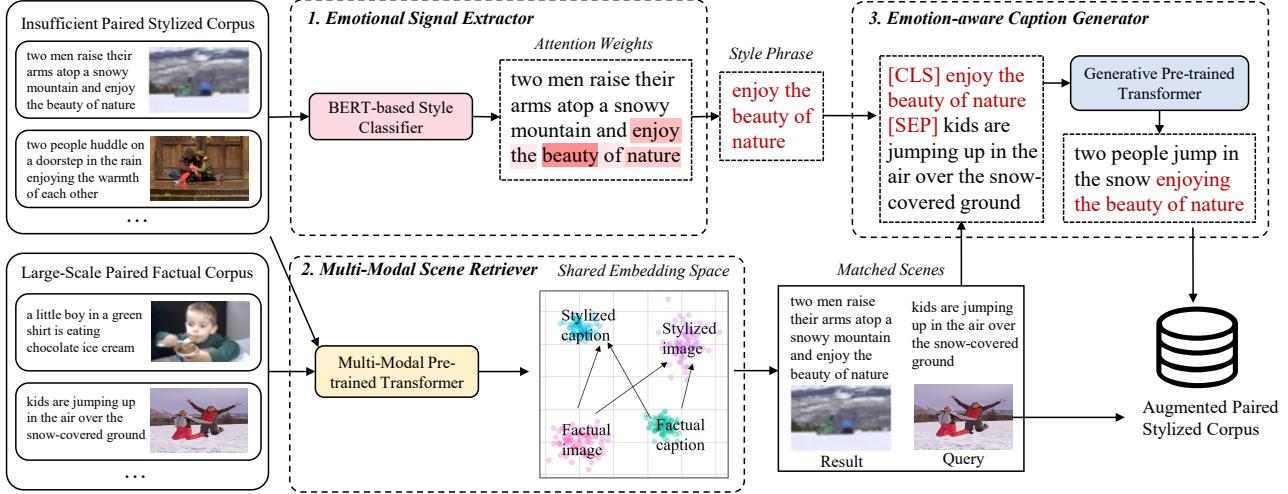


Figure 2: Overview of the proposed Extract-Retrieve-Generate data augmentation framework. First, the **Emotional Signal Extractor** extracts style phrases of the stylized sentences according to the attention weights. Second, for each factual image-caption pair, the **Multi-Modal Scene Retriever** retrieves the similar stylized image-caption pair in a shared embedding space through four types of retrieval method, i.e., image-to-image, image-to-text, text-to-image, and text-to-text retrieval. Finally, given the style phrase of the retrieved image-caption pair and the factual caption of the queried pair, the **Emotion-aware Caption Generator** generates a stylized caption for the image of the queried pair.

y_i^s denote the i -th image and its corresponding stylized caption, respectively, and a large-scale paired factual corpus $\mathcal{P} = \{(x_j^f, y_j^f)|_j\}$, where x_j^f and y_j^f denote the j -th image and its corresponding factual caption. The objective of our augmentation task is to produce large-scale paired stylized data \mathcal{D}_a , based on large-scale factual corpus \mathcal{P} and insufficient paired stylized data \mathcal{D}_s . With original paired stylized data \mathcal{D}_s and augmented data \mathcal{D}_a , we can train a better stylized image captioning model in a supervised way.

Figure 2 shows the whole pipeline of the proposed augmentation method. It consists of three components: 1) *The emotional signal extractor*, which aims to extract the style phrase of y^s . 2) *The multi-modal scene retriever*, which aims to retrieve a set of similar scenes in the small stylized dataset, given an image in the large factual dataset. 3) *The emotion-aware caption generator*, which leverages the style phrases of similar scenes retrieved above to generate stylized captions for images in the large-scale factual corpus \mathcal{P} . With the above components, we are able to form a new large-scale paired stylized data $\mathcal{D}_a = \{(x_k^f, \hat{y}_k^s)|_k\}$ to augment the original insufficient training set \mathcal{D}_s . In the following sections, we will describe these three components in detail.

3.2 Emotional Signal Extractor

Given an image x^s and its corresponding stylized caption y^s , the module is responsible for extracting the style phrase of the caption in an unsupervised way.

The structure of our extractor is inspired by [19] and [33], which recognize vulnerable words for text adversarial attack and attribute words for text style transfer by iteratively deleting words and observing the changes of output. Similarly, the style phrase of a stylized caption is often localized to certain words or phrases. A style

classifier will be confused about its style if we remove these words, leading to significantly changes of output.

With the above observation, our extractor is modeled as a well-trained style classifier, determining which word has a high impact on the final output. Formally, let $y^s = [w_0, \dots, w_t, \dots]$ denote the input stylized caption with style label s , and $I(w_t)$ is defined as the importance score of w_t to measure how much it contributes to the style of the caption. We rank all the words according to their importance scores in descending order and only extract ϵ proportion of the most important words as the style phrase. ϵ is also referred to as extraction proportion.

In practice, we adopt a BERT-based classifier [6] fine-tuned on the stylized captions as our extractor because of its strong performance on various NLP downstream tasks. The importance score is defined as the attention weights of other input tokens that correspond to $[CLS]$ token. Since BERT is stacked with multiple layers and multiple heads of transformer blocks, we can obtain multiple attention weights of words corresponding to every layer l and head h , formally:

$$I_{h,l}(w_t) = \text{softmax}_{w_t \in y^s} \left(\frac{Q_{h,l,[CLS]} K_{h,l,w_t}^T}{\sqrt{d_k}} \right), \quad (1)$$

where $Q_{h,l,[CLS]}$ and K_{h,l,w_t} are the query vector of $[CLS]$ and the key vector of w_t on the layer l and head h , and d_k is a scaling factor.

However, every layer and head encode different aspects of semantic and linguistic structure, we need to determine which attention weights are really related to the linguistic style aspect of captions. Concretely, for each layer and head, we remove ϵ percent of the most important words denoted as w_ϵ to form a new reduced caption $y_{/w_\epsilon}^s$. Following [33], we define the confidence score $c(h, l, y^s)$ for each layer and head to measure the changes of output probability,

Table 1: Four types of scene retrieval, i.e., image-to-image, image-to-text, text-to-image, and text-to-text retrieval.

Retrieval Type	Query	Target
image-to-image (i2i)	image x^f	image x^s with style phrase e^s
text-to-text (t2t)	text y^f	text y^s with style phrase e^s
image-to-text (i2t)	image x^f	text y^s with style phrase e^s
text-to-image (t2i)	text y^f	image x^s with style phrase e^s

formally:

$$c(h, l, y^s) = \frac{p(s|y_{/w_e}^s)}{\sum_{s' \in S'} p(s'|y_{/w_e}^s)}, \quad (2)$$

where $S' = S - \{s\}$ is the remaining style labels of the label set S . The lower the confidence score, the more confused the classifier is about its style, and the greater the change of output. Finally, we choose the layer l_s and head h_s that minimize the average confidence score of all stylized captions in \mathcal{D}_s to form the importance score $I_{h_s, l_s}(w_t)$.

For each stylized caption, measured by $I_{h_s, l_s}(w_t)$, we extract ϵ percent of the most important words in the caption y^s as the style phrase e^s . In this way, we update the paired stylized data \mathcal{D}_s to $\mathcal{D}'_s = \{(x_i^s, y_i^s, e_i^s)|_i\}$.

3.3 Multi-Modal Scene Retriever

To follow the intuition that similar scenes may be described in a similar stylized expression, the module is responsible for retrieving a set of style phrases from \mathcal{D}'_s for images in the corpus \mathcal{P} by utilizing multi-modal cues.

As shown in Figure 2, we adopt a multi-modal pre-trained transformer called OSCAR [20] as a unified multi-modal encoder $U(\cdot)$ to map all the texts and images into d -dimensional vectors in a shared embedding space.

Specifically, for each input image, we firstly extract K region-level visual features, with the Faster R-CNN [29] model pre-trained on Visual Genomes [17] as the input of pre-trained transformer. Each image and caption are passed separately through the pre-trained transformer model, which outputs the final hidden representation of [CLS] token as the respective image representation E_x and caption representation E_y . The similarity is defined as the cosine of the angle between their embedding vectors, formally:

$$\text{sim}(x, y) = \frac{E_x \cdot E_y}{\|E_x\| \|E_y\|} \quad (3)$$

Subsequently, we adopt an efficient similarity search tool called Faiss [15] to implement the retrieval process through four types of scene retrieval, as shown in Table 1.

In this way, we can retrieve more diverse style phrases for images in large-scale factual data \mathcal{P} .

3.4 Emotion-aware Caption Generator

With the retrieved style phrases, the caption generator module is responsible for generating stylized captions for images in the corpus \mathcal{P} to construct new paired stylized samples.

To achieve this, we regard the factual caption y^f in the corpus \mathcal{P} as the content of corresponding images and the retrieved style phrase e^s as the potential emotional prompt when describing the

image. The conditional probability of generating captions given above information is defined as follows:

$$p(\hat{y}^s|x^f) \approx p(\hat{y}^s|y^f, e^s) \quad (4)$$

Motivated by the recent remarkable performance of GPT [27], a powerful generative transformer pre-trained on large-scale unlabeled texts, we adopt GPT-2 as our generator and fine-tune it to learn the above conditional probability.

Let θ denote the parameters of GPT-2. In the fine-tuning process, for each stylized caption y^s in the \mathcal{D}_s , we first extract the style phrase e^s and regard the rest part $y_{/w_e}^s$ as the content of the image. Thus, the fine-tuning objective is to reconstruct the original stylized caption, i.e., minimizing the following objective:

$$L(\theta) = - \sum_{t=1}^T \log p(y_t^s|y_{1:t-1}^s; e^s; y_{/w_e}^s) \quad (5)$$

When inferring a new stylized caption, a sample of the process is illustrated as follows:

- (1) Retrieved style phrase: *enjoy the beauty of nature*
- (2) Factual caption in the corpus \mathcal{P} : *Kids are jumping up in the air over the snow-covered ground*
- (3) Input representation: [CLS] enjoy the beauty of nature [SEP]
Kids are jumping up in the air over the snow-covered ground
- (4) Output: two people jump in the snow enjoying the beauty of nature

Besides, we adopt the following three criteria to guarantee the quality of generated data: 1) the generated caption sentence can be classified by the BERT style classifier (see Sect. 4.4) as that target style, 2) its perplexity < 80 (see Sect. 4.4), 3) scene retrieval similarity score > 0.6 (see Fig.3). Those valid generated caption samples are used to construct paired augmented data.

4 EXPERIMENTS

4.1 Experimental Datasets

To validate and analyze the effectiveness of our data augmentation framework on the stylized image caption task, we conduct our experiments on two popular stylized image caption datasets: FlickrStyle10K and SentiCap.

FlickrStyle10K [8] is collected and built on Flickr30K [13] image caption dataset. The original FlickrStyle10K dataset has 10,000 pairs of images and stylized captions including humorous and romantic styles. However, only 7,000 pairs from official training set is now publicly accessible. We take the same processing method following Guo et al.[11] and Zhao et al.[47] that randomly sample 6,000 images as our insufficient training data and the rest images are used for testing. To obtain the optimal setting of the hyper-parameters, we randomly split 10% samples of the training set as the validation set.

SentiCap [24] constructs an additional set of captions with positive and negative sentiments using images from the MSCOCO [21] validation partition. 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 as the validation set.

In all experiments, we construct the large-scale paired factual corpus with the MSCOCO training partition and Flickr30K dataset

Table 2: Performance comparisons on the test splits of four styles. “Original” and “Augmentation” mean the models use the original or augmented training data. B-n, M, C, cls., and ppl. are short for BLEU-n, METEOR, CIDEr, style classification accuracy, and perplexity, respectively. ΔC and $\Delta \text{cls.}$ measure the performance gains in terms of CIDEr and cls. metrics.

Style	Model	Original						Augmentation						$\Delta C(\%)$	$\Delta \text{cls.}(\%)$
		B-1	B-3	M	C	cls.	ppl.	B-1	B-3	M	C	cls.	ppl.		
Humor	NIC	27.3	8.3	10.9	39.2	82.8	6.0	28.4	9.6	11.8	45.2	97.5	9.2	+6.0	+14.7
	Up-Down	29.9	9.9	12.2	50.5	80.7	8.6	30.3	10.7	13.3	54.2	98.6	8.7	+3.7	+17.9
	StyleNet	24.6	7.3	10.3	34.7	42.5	7.3	24.6	6.8	9.8	28.3	96.2	7.6	-6.4	+53.7
	SAN	27.6	8.1	11.2	38.7	87.8	8.4	29.5	9.9	12.5	47.2	99.4	13.7	+8.5	+11.6
	VinVL	26.5	7.6	11.3	38.2	89.1	12.9	28.9	9.0	12.2	44.4	99.7	12.2	+6.2	+10.6
Roman	NIC	28.1	8.7	11.2	41.6	86.5	5.4	31.0	9.9	12.4	47.2	97.5	8.0	+5.6	+11.0
	Up-Down	30.7	10.7	12.9	55.5	86.4	7.7	32.8	11.5	13.6	57.9	97.4	8.6	+2.4	+11.0
	StyleNet	24.8	7.4	10.3	34.4	57.1	6.9	25.8	7.0	10.2	30.3	92.7	5.9	-4.1	+35.6
	SAN	28.3	8.7	11.5	42.1	90.9	9.1	30.9	10.9	13.0	53.3	99.6	13.1	+11.2	+8.7
	VinVL	28.7	8.3	11.8	41.2	96.6	10.1	31.1	10.1	13.1	52.1	96.8	11.0	+10.9	+0.20
Pos	NIC	40.3	15.2	11.8	29.9	99.4	5.2	50.5	22.6	17.3	69.1	100.0	12.2	+39.2	+6.0
	Up-Down	49.6	20.7	15.9	52.3	92.8	9.8	52.4	24.4	18.1	77.7	100.0	10.3	+25.4	+7.2
	StyleNet	43.9	13.0	11.6	30.8	0.0	10.3	33.2	8.7	8.3	14.9	100.0	4.9	-15.9	+100.0
	SAN	44.0	17.2	13.5	41.5	93.2	6.4	53.0	23.4	18.1	72.0	100.0	11.7	+30.5	+6.8
	VinVL	45.6	15.6	14.4	43.3	76.9	12.6	53.4	23.4	18.0	75.9	100.0	12.4	+32.6	+23.1
Neg	NIC	44.5	15.7	14.2	33.2	96.0	5.8	51.4	19.9	17.4	62.2	100.0	12.5	+29.0	+4.0
	Up-Down	51.0	19.6	16.5	50.8	91.7	11.3	52.6	21.6	18.0	68.3	100.0	8.5	+17.5	+8.3
	StyleNet	43.5	12.9	11.7	27.8	10.1	8.5	36.3	8.2	10.6	13.6	100.0	9.0	-14.2	+89.9
	SAN	44.6	15.2	14.0	36.0	92.3	8.0	51.2	20.5	17.6	67.0	100.0	14.8	+31.0	+7.7
	VinVL	44.6	14.6	15.5	41.4	88.1	15.1	53.1	21.2	18.6	70.0	100.0	13.7	+28.6	+11.9

excluding images and corresponding factual captions in the above stylized image caption datasets.

4.2 Experimental Models

To validate the effectiveness and applicability of our method, we implement the proposed data augmentation framework on following representative models, which can be classified into the three categories: supervised models (NIC and Up-Down), unsupervised models (StyleNet), and Transformer-based models (SAN and VinVL).

- *Neural Image Caption (NIC)* [37] is a standard encoder-decoder image caption pipeline, which encodes the image using CNN and decodes it using LSTM.
- *Up-Down* [2] employs a bottom-up and top-down attention mechanism with the two-layer LSTM, which enables attention to be calculated more naturally at the level of objects and other salient regions.
- *StyleNet* [8] devises a model component, named factored LSTM, which automatically distills the style factors in the monolingual text corpus. This model can explicitly control the style in the caption generation process so as to produce attractive visual captions with the desired style.
- *Self-Attention Network (SAN)* [12] consists of an image encoder and a caption decoder, both of which are composed of a stack of transformer blocks. The inputs to the encoder are the region-based visual features and the decoder takes the attended visual features and the embeddings of the previous words to predict the next word recursively.

- *VinVL* [45] is a powerful multi-modal pre-trained model for visual-language tasks, which is pre-trained on a large-scale text-image corpora and achieves new state-of-the-art results on several public benchmarks include image captioning.

4.3 Comparison Models

We also compare our approach with several previous state-of-the-art methods for stylized image captioning.

- Style-Factual LSTM (SF-LSTM) [5] proposes an adaptive learning approach based on a reference factual model and provides factual knowledge to the model as the model learns from stylized caption labels, and can adaptively compute how much information to supply at each time step.
- MSCap [11] proposes an adversarial learning framework with several well-designed modules to generate visually-grounded and style-controllable captions, which is trained on unpaired stylized corpus.
- MemCap [47] is a novel stylized image captioning method that explicitly encodes the knowledge about linguistic styles with memory mechanism, which is also trained on unpaired stylized corpus.

4.4 Evaluation Metrics

We evaluate our method on several baseline models from two aspects: the ability of generating fluent and accurate captions, and the performance of generating captions in target linguistic style. Following the related baseline work [11, 47], we measure the sentence

fluency and relevancy using the metrics of BLEU [26] (including BLEU-1 and BLEU-3), METEOR [3], CIDEr [35].

In addition, we also measure the sentence stylishness using the style classification accuracy (cls.) and the average perplexity (ppl.) following [11] and [47]. Accuracy of style classification is defined as the percentage of sentences that correctly reflect the desired styles. For that purpose, we use BERT to train a style classifier for each of four styles with stylized sentences from FlickrStyle10K and SentiCap datasets and factual sentences from MSCOCO dataset. These classifiers achieve accuracies of 96.5%, 97.4%, 99.3%, and 98.8% on the test sets of humorous, romantic, negative, and positive styles respectively. As for the average perplexity, we use the SRILM toolkit to train a tri-gram based statistical language model for each of four styles on corresponding stylized sentences. The language models achieve perplexities of 76, 61, 47, and 52 on the test sets of four styles respectively. A lower score indicates that the generated sentences are more fluent and better reflect the desired linguistic style.

5 RESULTS AND ANALYSIS

5.1 Quantitative Analysis

Main Results. We report the results of several experimental models trained with the augmented and original data in Table 2. As expected, all models trained with the augmented data substantially outperform those with the original data. Even for the traditional NIC model, both sentence stylishness (measured by cls.) and sentence relevancy (measured by BLEU, CIDEr, and METEOR) are greatly improved compared to the model trained on the original data. For example, under the evaluation of cls. and CIDEr, NIC achieves 6% improvement and 14.7% improvement on humorous style. Benefiting from the transformer architecture and pre-training technology, SAN and VinVL achieve nearly the best results. For example, VinVL achieves the best scores of all metrics on negative style compared with other models except for perplexity (ppl.). As for the perplexity metric, we claim that it does not correlate entirely well with stylishness, since the average perplexities on the test set of four styles are much higher than the results of Table 2.

However, a seemingly abnormal observation is that StyleNet trained on the augmented data leads to a performance drop with respect to sentence relevancy but a great performance increase on sentence stylishness. We attribute it to its special training process, i.e., cross-training on a large number of factual captions and a small number of style captions. The imbalance of data leads to an excessive bias towards factual information and neglect of style expression. With augmented data, it can better balance the two kinds of information, although at the expense of some performance related to sentence relevance. We also observe that VinVL trained on the original data has achieved 96.6 on the evaluation of cls. for the romantic style. There is only 0.2% improvement with augmented data. Combined with the result of CIDEr, it is obvious that VinVL trained on the original romantic data prefers to improve sentence stylishness over sentence relevance.

From the above observations, we conclude that the key to this task is how to make a good trade-off between sentence relevance and sentence stylishness. It is empirically confirmed that our augmentation approach can generate more stylized samples for doing

Table 3: Performance comparison with state-of-the-art methods. “Ours(SAN)” uses the augmented data for training.

Style	Model	B-1	B-3	M	C	cls.	ppl.
Humor	SF-LSTM	27.4	8.5	11.0	39.5	-	-
	MSCap	16.3	1.9	5.3	15.2	91.3	22.7
	MemCap	19.9	4.3	7.4	19.4	98.9	16.4
	Ours (SAN)	29.5	9.9	12.5	47.2	99.4	13.7
Roman	SF-LSTM	27.8	8.2	11.2	37.5	-	-
	MSCap	17.0	2.0	5.4	10.1	88.7	20.4
	MemCap	21.2	4.8	8.4	22.4	98.7	14.4
	Ours (SAN)	30.9	10.9	13.0	53.3	99.6	13.1
Pos	SF-LSTM	50.5	19.1	16.6	60.0	-	-
	MSCap	46.9	16.2	16.8	55.3	92.5	19.6
	MemCap	50.8	17.1	16.6	54.4	99.8	13.0
	Ours (SAN)	53.0	23.4	18.1	72.0	100.0	11.7
Neg	SF-LSTM	50.3	20.1	16.2	59.7	-	-
	MSCap	45.5	15.4	16.2	51.6	93.4	19.2
	MemCap	48.7	19.6	15.8	60.6	93.1	14.6
	Ours (SAN)	51.2	20.5	17.6	67.0	100.0	14.8

Table 4: Visualizations of the attention weights of stylized sentences produced by the proposed extractor.

Style	Stylized Sentences
Humor	a boy wearing an orange coat winks at the camera suspiciously
Roman	a young boy in a red shirt plays on a mini trampoline in a grassy field having fun with some friends
Pos	a pretty woman smiling on her favorite street
Neg	a bad image of a person in a poor city holding out stop sign near a child

a better trade-off between the above aspects and training a more powerful model.

Comparison with State-of-the-Arts. We select SAN with augmented data to compare with several state-of-the-art methods, i.e., SF-LSTM, MSCap, and MemCap. Table 3 shows that SAN outperforms previous state-of-the-art models by a significantly large margin. In addition, regardless of the effect of the transformer architecture, NIC trained with augmented data also outperforms them across most metrics, as shown in Table 2. The above results validate the effectiveness of our data augmentation framework. Meanwhile, those previous models can also be trained on our augmented data to obtain greater improvements without any architecture changes.

Compare MemCap/MSCap with baselines trained with original data, another seemingly abnormal observation is that there is a huge gap in sentence relevance. We claim that these well-designed models tend to sacrifice certain sentence relevance and improve the sentence stylishness. Baseline models trained simply with insufficient parallel data have difficulties in capturing stylishness and tend to improve relevance. Our ERG data augmentation framework is able to make a good trade-off between two aspects.

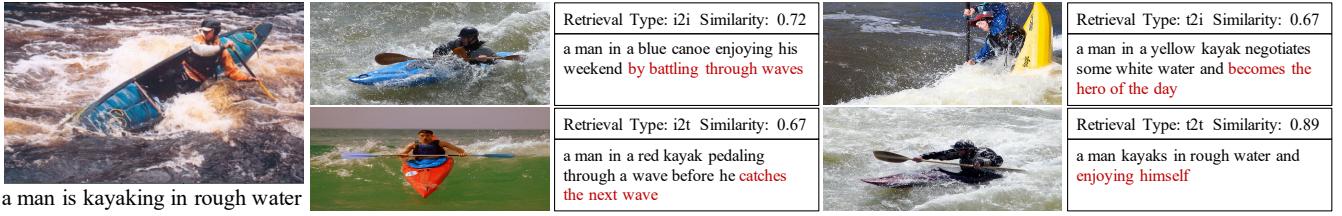


Figure 3: Retrieved stylized image-caption pairs with four types of retrieval. The left side is the queried image-caption factual pair. Words in red represent the style phrase. The similarity means the cosine similarity between the query and the result.



Figure 4: Augmentation examples generated by our framework. Words in purple, green, red, and blue represent the retrieved humorous, romantic, positive, and negative style phrases, respectively. Although there are some noise in the retrieved style parts, our generator can refine them correctly, e.g., *each tight* → *each other tight*.

Table 5: Effect of Different Retrieval Types.

Retrieval Type	#Samples	B-3	M	C	cls.
image-to-image (i2i)	7087	9.55	11.87	46.03	94.1
image-to-text (i2t)	3955	9.36	11.76	44.88	91.4
text-to-image (t2i)	1382	8.96	11.49	44.73	89.7
text-to-text (t2t)	29116	9.73	12.18	46.73	96.9
Ours (all)	37699	9.89	12.44	47.17	97.5

Table 6: Effect of Different Caption Generators.

Style	Generator	B-3	M	C	cls.	ppl.
Humor	T5 (SEQ2SEQ)	9.25	11.86	43.09	98.9	9.9
	GPT-2 (LM)	9.57	11.76	45.21	97.5	9.2
Roman	T5 (SEQ2SEQ)	9.93	12.38	49.01	96.4	8.2
	GPT-2 (LM)	9.89	12.44	47.17	97.5	8.0

5.2 Qualitative Analysis

We visualize the heatmap of words from the stylized sentences and their attention weights from the extractor corresponding to [CLS] token, as illustrated in Table 4. According to the heatmap, whether each word belongs to the style phrase can be seen intuitively. It is obvious that attention weights of sentences highly correlate with sentence stylishness. It confirms that our extractor is able to effectively discern the style phrase of the sentence.

As shown in Figure 3, we also presents a typical retrieval result of our multi-modal scene retriever. From the retrieval results, we can see our retriever can find more stylized matching pairs with low noise. In this way, we can add more diverse styles to the factual captions in the large-scale factual corpus.

Figure 4 presents some augmented examples to illustrate the effectiveness of our augmentation approach intuitively. Our framework can generate grammatically correct and diversified captions for scenes in the large factual dataset, even if style phrases extracted and retrieved from the original stylized data are slightly noisy.

5.3 Ablation Studies

We further conduct experiments to validate the effectiveness of different components. Unless otherwise specified, all the experiments are carried out using the simplest NIC model with augmented data on romantic style.

Effect of the Size of Augmentation Data. By training the NIC model on the 10%, 30%, 50%, 70%, and 100% augmentation data, We investigate the impact of different-sized augmentation data. Specifically, the augmentation data are first sorted in ascending order of perplexity (see Sect. 4.4). Then, the top 10%, 30%, 50%, 70%, 100% of sorted augmented data are selected to experiment. As shown in Figure 5b, we observe that the score of cls. increases as the size of augmentation data increases. When data augment to a certain size, the metric stops increasing. The phenomenon indicates the performance of this task can be boosted by increasing size of training data, but there is an upper bound of learning effective stylized knowledge. Meanwhile, there is a slight performance drop in terms of sentence relevance (from CIDEr of 48.50 to 47.17), since augmented data of larger size have more noise samples.

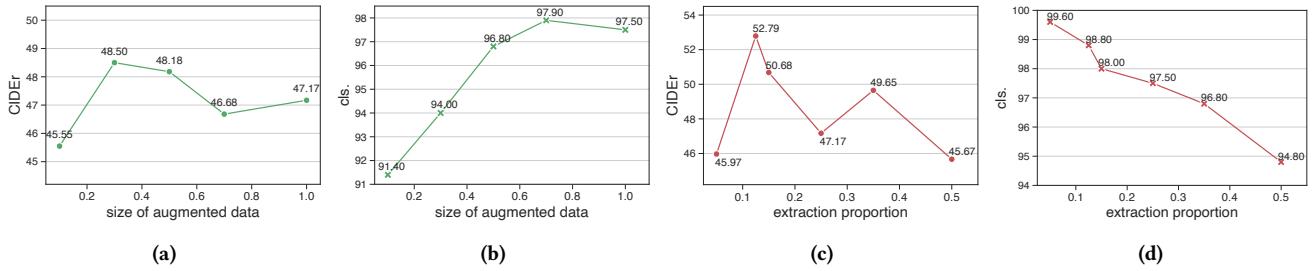


Figure 5: (a) CIDEr score with different size of augmented data. (b) cls. score with different size of augmented data. (c) CIDEr score with different extraction proportion. (d) cls. score with different extraction proportion.

Effect of Extraction Proportion. linguistic style is often localized to a small subset of words of a sentence. We investigate the impact of extraction proportion ϵ of the words ranked by the attention weights. Figure 5d shows that the score of the cls. metric gradually decreases as the proportion increases, which indicates that a larger proportion will introduce more noise. Meanwhile, Figure 5c shows that higher or lower proportions will degrade scores of sentence relevance to a certain extent. Nevertheless, NIC still achieves CIDEr of 45.97 and cls. of 94.80 with the proportion of 0.05 and 0.5, respectively, empirically confirming that our framework is insensitive to choice of the proportion within a reasonable range.

Effect of Different Retrieval Types. For validating the effectiveness of the multi-modal scene retriever, we also test the effects of different retrieval methods. Table 5 reports performance scores of our approach with i2i, i2t, t2i and t2t retrieval respectively. From Table 5, we conclude that using multiple retrieval methods at the same time can get more augmented samples and improve the performance of the downstream task, in terms of cls. and CIDEr. From the retrieval cases in Figure 3, it is more intuitively seen that our retriever can find more stylized matching pairs with low noise.

Effect of Different Caption Generators. To present the flexibility of our augmentation framework, we analyze the impact of different state-of-the-art generative models (GPT-2 and T5) as the caption generator. While GPT-2 is a novel and powerful pre-trained language model, T5 [28] is another pre-trained model which adopts sequence-to-sequence architecture. As shown in Table 6, the pre-trained models of two architectures have no obvious differences in downstream tasks, and both achieve good performance scores. From the above results, we conclude that our framework is powerful and flexible, which fully leverages the knowledge from large pre-trained transformers in a plug-and-play manner.

6 DISCUSSION

In this section, we discuss the necessity of parallel data and the robustness of our framework, which are essential for validating the effectiveness and generalization of our proposed method.

The unneccessity of parallel data. The purpose of this paper is to augment the insufficient paired stylized data. However, if there are no stylized captions with aligned images at the beginning, our framework can also run well with unpaired stylized text and existing large-scale factual image-caption datasets. First, the process of extracting style phrases from the stylized text does not require

aligned images. Our proposed emotional signal extractor just processes the stylized text to extract style phrases without requiring aligned images as the input. Second, when implementing the retriever with the text-to-text (t2t) or image-to-text (i2t) retrieval method, there are also no requirements for aligned images. From the results of Table 5, we observe that only using text-to-text retrieval method can also achieve competitive results. Therefore, our framework can generate stylized paired data only with unpaired stylized text and existing large-scale factual image-caption pairs.

The robustness of our framework. Our framework consists of three separated modules running as an extract-retrieve-generate pipeline. Although the noise may be introduced in each process, our framework is still robust. From the augmented examples in Figure 4, our generator can produce grammatically correct sentences, even if there are some mistakes in the retrieved styles. For example, it discerns a missing word *other* in “while holding each tight”, and modifies “*a cat*” in the beginning of original sentence with “*an annoying cat*”. Meanwhile, the results of various experimental models in this task confirm that our framework is effective and robust.

7 CONCLUSION

In this paper, we propose a novel extract-retrieve-generate data augmentation framework for stylized image captioning. The proposed framework first extracts the style phrases from existing stylized sentences. Then the framework further retrieves a set of similar scenes through a plug-able multi-modal scene retriever. With the style phrases in the stylized captions about similar scenes, the framework generates corresponding stylized captions for images in the large-scale factual corpus. Results of extensive experiments show that the framework can alleviate the data scarcity problem effectively. It also significantly boosts the performance of several existing image captioning models in both supervised and unsupervised settings, which outperforms the state-of-the-art stylized image captioning methods by a substantial margin.

ACKNOWLEDGMENTS

We thank Tong Li, Yongliang Shen, Tian Lan, Chenyu Yan, Yan Yu, and Qinghao Ye for useful help. This work was supported by the NSFC projects (No. 62072399, No. U19B2042), Chinese Knowledge Center for Engineering Sciences and Technology, MoE Engineering Research Center of Digital Library, Alibaba-Zhejiang University Joint Research Institute of Frontier Technologies, and the Fundamental Research Funds for the Central Universities.

REFERENCES

- [1] Ethel Mary Abernethy. 1940. The effect of changed environmental conditions upon the results of college examinations. *The Journal of Psychology* 10, 2 (1940), 293–301.
- [2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018*. IEEE Computer Society, 6077–6086. <https://doi.org/10.1109/CVPR.2018.00636>
- [3] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare R. Voss (Eds.). Association for Computational Linguistics, 65–72. <https://www.aclweb.org/anthology/W05-0909/>
- [4] Cheng-Kuan Chen, Zhufeng Pan, Ming-Yu Liu, and Min Sun. 2019. Unsupervised stylish image description generation via domain layer norm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 8151–8158.
- [5] Tianlang Chen, Zhongping Zhang, Quanzeng You, Chen Fang, Zhaowen Wang, Hailin Jin, and Jiebo Luo. 2018. “Factual” or “Emotional”: Stylistized Image Captioning with Adaptive Learning and Attention. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 519–535.
- [6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, 2019, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 4171–4186. <https://doi.org/10.18653/v1/n19-1423>
- [7] Stanley G Dulsky. 1935. The effect of a change of background on recall and relearning. *Journal of Experimental Psychology* 18, 6 (1935), 725.
- [8] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017. Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3137–3146.
- [9] Duncan R Godden and Alan D Baddeley. 1975. Context-dependent memory in two natural environments: On land and underwater. *British Journal of psychology* 66, 3 (1975), 325–331.
- [10] Longteng Guo, Jing Liu, Jinhui Tang, Jiangwei Li, Wei Luo, and Hanqing Lu. 2019. Aligning Linguistic Words and Visual Semantic Units for Image Captioning. In *Proceedings of the 27th ACM International Conference on Multimedia (Nice, France) (MM ’19)*. Association for Computing Machinery, New York, NY, USA, 765–773. <https://doi.org/10.1145/3343031.3350943>
- [11] Longteng Guo, Jing Liu, Peng Yao, Jiangwei Li, and Hanqing Lu. 2019. MSCap: Multi-style image captioning with unpaired stylized text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4204–4213.
- [12] Longteng Guo, Jing Liu, Xinxin Zhu, Peng Yao, Shichen Lu, and Hanqing Lu. 2020. Normalized and Geometry-Aware Self-Attention Network for Image Captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13–19, 2020*. IEEE, 10324–10333. <https://doi.org/10.1109/CVPR42600.2020.01034>
- [13] Mical Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research* 47 (2013), 853–899.
- [14] MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CSUR)* 51, 6 (2019), 1–36.
- [15] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with GPUs. *arXiv preprint arXiv:1702.08734* (2017).
- [16] Sulabh Katiyar and Samir Kumar Borgohain. 2021. Image Captioning using Deep Stacked LSTMs, Contextual Word Embeddings and Data Augmentation. *CoRR* abs/2102.11237 (2021). <https://arxiv.org/abs/2102.11237>
- [17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *Int. J. Comput. Vis.* 123, 1 (2017), 32–73.
- [18] Kun Li, Chengbo Chen, Xiaojuan Quan, Qing Ling, and Yan Song. 2020. Conditional Augmentation for Aspect Term Extraction via Masked Sequence-to-Sequence Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5–10, 2020*. Association for Computational Linguistics, 7056–7066.
- [19] Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. BERT-ATTACK: Adversarial Attack Against BERT Using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16–20, 2020*. Association for Computational Linguistics, 6193–6202. <https://doi.org/10.18653/v1/2020.emnlp-main.500>
- [20] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*. 121–137.
- [21] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V*. 740–755.
- [22] Canjie Luo, Yuanzhi Zhu, Lianwen Jin, and Yongpan Wang. 2020. Learn to Augment: Joint Data Augmentation and Network Optimization for Text Recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13–19, 2020*. IEEE, 13743–13752.
- [23] Alexander Mathews, Lexing Xie, and Xuming He. 2018. Semstyle: Learning to generate stylised image captions using unaligned text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8591–8600.
- [24] Alexander Patrick Mathews, Lexing Xie, and Xuming He. 2016. Senticap: Generating image descriptions with sentiments. In *Thirtyfirst AAAI conference on artificial intelligence*.
- [25] Omid Mohamad Nezami, Mark Dras, Stephen Wan, Cécile Paris, and Len Hamey. 2019. Towards Generating Stylized Image Captions via Adversarial Training. In *Pacific Rim International Conference on Artificial Intelligence*. Springer, 270–284.
- [26] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6–12, 2002, Philadelphia, PA, USA*. 311–318. <http://www.aclweb.org/anthology/P02-1040.pdf>
- [27] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.
- [28] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *CoRR* abs/1910.10683 (2019). <https://arxiv.org/abs/1910.10683>
- [29] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7–12, 2015, Montreal, Quebec, Canada*. 91–99.
- [30] Shurong Sheng and Marie-Francine Moens. 2019. Generating Captions for Images of Ancient Artworks. In *Proceedings of the 27th ACM International Conference on Multimedia*, 2478–2486.
- [31] Kurt Shuster, Samuel Humeau, Hexiang Hu, Antoine Bordes, and Jason Weston. 2019. Engaging image captioning via personality. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 12516–12526.
- [32] Yuqiang Song, Shizhe Chen, Yida Zhao, and Qin Jin. 2019. Unpaired Cross-Lingual Image Caption Generation with Self-Supervised Rewards. In *Proceedings of the 27th ACM International Conference on Multimedia (Nice, France) (MM ’19)*. Association for Computing Machinery, New York, NY, USA, 784–792. <https://doi.org/10.1145/3343031.3350996>
- [33] Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheshwaran. 2019. “Transforming” Delete, Retrieve, Generate Approach for Controlled Text Style Transfer. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3–7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojuan Wan (Eds.). Association for Computational Linguistics, 3267–3277. <https://doi.org/10.18653/v1/D19-1322>
- [34] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*.
- [35] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.
- [36] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3156–3164.
- [37] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7–12, 2015*. IEEE Computer Society, 3156–3164. <https://doi.org/10.1109/CVPR.2015.7298935>
- [38] Cheng Wang, Haojin Yang, and Christoph Meinel. 2018. Image Captioning with Deep Bidirectional LSTMs and Multi-Task Learning. *ACM Trans. Multim. Comput. Commun. Appl.* 14, 2s (2018), 40:1–40:20.
- [39] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. 2015. Show, attend and tell: neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37*. 2048–2057.

- [40] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. 2017. Boosting image captioning with attributes. In *Proceedings of the IEEE International Conference on Computer Vision*. 4894–4902.
- [41] Xingxu Yao, Dongyu She, Sicheng Zhao, Jie Liang, Yu-Kun Lai, and Jufeng Yang. 2019. Attention-Aware Polarity Sensitive Embedding for Affective Image Retrieval. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), 1140–1150.
- [42] Guojun Yin, Lu Sheng, Bin Liu, Nenghai Yu, Xiaogang Wang, and Jing Shao. 2019. Context and Attribute Grounded Dense Captioning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), 6234–6243.
- [43] Quanzeng You, Hailin Jin, and Jiebo Luo. 2018. Image captioning at will: A versatile scheme for effectively injecting sentiments into image descriptions. *arXiv preprint arXiv:1801.10121* (2018).
- [44] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4651–4659.
- [45] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. VinVL: Making Visual Representations Matter in Vision-Language Models. *CoRR abs/2101.00529* (2021). arXiv:2101.00529 <https://arxiv.org/abs/2101.00529>
- [46] Yumeng Zhang, Gaoguo Jia, Li Chen, Mingrui Zhang, and Junhai Yong. 2020. Self-Paced Video Data Augmentation by Generative Adversarial Networks with Insufficient Samples. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*. ACM, 1652–1660.
- [47] Wentian Zhao, Xinxiao Wu, and Xiaoxun Zhang. 2020. MemCap: Memorizing Style Knowledge for Image Captioning. In *Thirty-Fourth AAAI conference on artificial intelligence*.

APPENDICES

A. Augmentation Process

Given an insufficient paired stylized data \mathcal{D}_s and a large-scale paired factual data \mathcal{P} , we can obtain the augmented large-scale paired stylized data \mathcal{D}_a through our Extract-Retrieve-Generate pipeline. Algorithm 1 shows the whole augmentation process.

Algorithm 1 Augmentation Process

Input:

\mathcal{D}_s : an insufficient paired stylized image-caption data.

\mathcal{P} : a large-scale paired factual image-caption corpus.

K : the number of types of retrieval ($K = 4$ means all scene retrievals in Table 1).

ϵ : the proportion of the most important words.

Output:

\mathcal{D}_a : a new large-scale paired stylized augmented data.

1: **procedure** EXTRACT(\mathcal{D}_s, ϵ)

2: $\mathcal{D}'_s \leftarrow \emptyset$

3: **for** (x^s, y^s) in \mathcal{D}_s **do**

4: $e^s \leftarrow \text{Extractor}(y^s, \epsilon)$

5: $\mathcal{D}'_s \leftarrow \mathcal{D}'_s \cup \{(x^s, y^s, e^s)\}$

6: **end**

7: **return** \mathcal{D}'_s

8: **end procedure**

9: **procedure** RETRIEVE($\mathcal{D}'_s, \mathcal{P}, K$)

10: $\mathcal{P}' \leftarrow \emptyset$

11: **for** (x^f, y^f) in \mathcal{P} **do**

12: $\{(x_i^s, y_i^s, e_i^s)|_{i=1}^K\} \leftarrow \text{Retriever}(x^f, y^f)$

13: $\mathcal{P}' \leftarrow \mathcal{P}' \cup \{(x_i^s, y_i^s, e_i^s, x^f, y^f)|_{i=1}^K\}$

14: **end**

15: **return** \mathcal{P}'

16: **end procedure**

17: **procedure** GENERATE(\mathcal{P}')

18: $\mathcal{D}_a \leftarrow \emptyset$

19: **for** $(x^s, y^s, e^s, x^f, y^f)$ in \mathcal{P}' **do**

20: $\hat{y}^s \leftarrow \text{Generator}(y^f, e^s)$

21: $\mathcal{D}_a \leftarrow \mathcal{D}_a \cup \{(x^f, \hat{y}^s)\}$

22: **end**

23: **return** \mathcal{D}_a

24: **end procedure**

25: $\mathcal{D}'_s \leftarrow \text{Extract}(\mathcal{D}_s, \epsilon);$

26: $\mathcal{P}' \leftarrow \text{Retrieve}(\mathcal{D}'_s, \mathcal{P}, K);$

27: $\mathcal{D}_a \leftarrow \text{Generate}(\mathcal{P}');$

28: **return** $\mathcal{D}_a;$

B. Implementation Details

Data Augmentation. For emotional signal extractor, we employ the pre-trained BERT-base model with the uncased version of tokenizer. We use a mini-batch size of 32 for romantic and humor styles and a mini-batch size of 64 for positive and negative styles. The learning rates of AdamW optimizers are set to 1e-5 and 2e-5, respectively. The numbers of fine-tuning epochs are both set to 3. The proportion of the most important words ϵ is set to 0.25. For multi-modal cross retriever, we use the pre-trained OSCAR as a unified multi-modal encoder to map texts and images into 768-dimensional vectors. For caption generator with emotional prompt, we employ the pre-trained GPT-2 model. The learning rate of AdamW optimizer is set to 5e-5 and the number of fine-tuning epochs is set to 1.

Experimental Models Settings. For each image, we employ the Faster-RCNN [29] detector with ResNet-101 to extract the top $K = 36$ region proposals and obtain a 2048-dimensional feature for each region. In all experimental models except for SAN and VinVL, the hidden size of decoder is set to 512, which is also the size of word embeddings. As for the training process, we train these models under cross-entropy loss with a mini-batch size of 64, and Adam optimizer is used with a learning rate of 5e-4. We increase the probability of feeding back a sample of the word posterior by 0.05 every 5 epochs. In particular, for StyleNet, we take the same hyper-parameters and training methods suggested in [8] to keep the original unsupervised settings. For SAN and VinVL, we also follow the hyper-parameters and settings suggested in [12] and [45].