symmetry

MDPI

Article

Better Understanding: Stylized Image Captioning with Style Attention and Adversarial Training

Zhenyu Yang [1,2,*], Qiao Liu [2] and Guojing Liu [2]

[1] School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China

[2] School of Computer Science and Technology, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250353, China; 1043118413@stu.qlu.edu.cn (Q.L.); 1043118421@stu.qlu.edu.cn (G.L.)

* Correspondence: yzy@qlu.edu.cn

Received: 19 October 2020 ; Accepted: 27 November 2020; Published: 30 November 2020

Abstract: Compared with traditional image captioning technology, stylized image captioning has broader application scenarios, such as a better understanding of images. However, stylized image captioning faces many challenges, the most important of which is how to make the model take into account both the image meta information and the style factor of the generated captions. In this paper, we propose a novel end-to-end stylized image captioning framework (ST-BR). Specifically, we first use a style transformer to model the factual information of images, and the style attention module learns style factor form a multi-style corpus, it is a symmetric structure on the whole. At the same time, we use back-reinforcement to evaluate the degree of consistency between the generated stylized captions with the image knowledge and specified style, respectively. These two parts further enhance the learning ability of the model through adversarial learning. Our experiment has achieved effective performance on the benchmark dataset.

Keywords: image understanding; stylized image captioning; symmetric end-to-end structure; style attention; back-reinforcement module

1. Introduction

In recent years, image captioning [1–4] has achieved great performance. However, the lack of personality and style knowledge limits its application prospects. For example, in the visual question and answer chat robots, stylized descriptions will make robots interact better with humans. Besides, generated medical reports can help doctors to make a more correct diagnosis. Jing et al. [5] utilized the method of deep learning to generate medical reports.

The task of stylized image captioning [6,7] is derived from traditional image captioning. It incorporates specific style factors (such as humor, romance, etc.) into captions to enable readers better understand images. Stylized image captioning has many application scenarios. For instance, generated stylized descriptions of products can attract more purchase behavior; image captions of multi-style will make visually impaired people perceive images from different perspectives. Zhang et al. [8] proposed an enlightening work. Their work incorporates detailed knowledge of the user's writing style to generate captions. The captions were shared and can receive more attention on social media.

Stylized image captioning is a branch of the image captioning. It is a challenging problem of balancing the factual information of images and specific style knowledge simultaneously.

Because stylized image captioning has more broad application prospects, some scholars have begun to pay attention to it and study. Gan et al. [6] used the method of knowledge distillation to extract the style factors in the corpus. Their work pays more attention to style and despises image information. Guo et al. [9] use multi-module feedback to generate stylized captions, which is consistent

with Gan et al. Mathews et al. [7] use a model to generate single-style captions and lack attention to multiple styles.

The pre-trained and fine-tuning stage constitute the learning process of stylized image captioning. The first stage makes the model have the ability to describe images by training in a factual dataset. However, the existing model simply describes the image information and does not fully use its factual information. In the second stage, the model generates captions in a specific writing style by training in the unpaired stylized corpus. However, most existing methods can only train one style corpus at a time.

Drawing on some recent work on traditional image captioning, we propose a novel transformer structure. The structure integrates a multi-style module based on the transformer [10] and can learn style factors from a multi-style corpus and memorize each style in training. Simultaneously, the transformer can model the image regions and the relationships between high-level and low-level information. In this way, the model can better integrate image information into the process of caption generation.

The existing stylized models generate captions which cannot effectively reflect the image's factual information. Taking romantic captions as an example, we can find the difference between the two in Figure 1. Romantic captions pay much attention to the subjective immersive experience based on the description information, while the factual caption is simply a combination of the factual information (objects, attributes, relationships, etc.) of the image. However, because romantic captions pay too much attention to style, they lose some factual information. It is easy to imagine that although image feature extraction and caption generation are two tasks, they are closely related. Considering this issue, based on the idea of adversarial, we propose a back-reinforcement module to evaluate whether the generated captions are accurate enough. In detail, we use factual knowledge discriminator and style factor discriminator, respectively, to carry on the adversarial training with caption generator. The method enables the model to strengthen the learning of image information and style factors.

Factual Caption : A collection of vases are on display in them.
Romantic Caption : I've been in a collection of vases on display.

(a)

Factual Caption : A living room with a couch and chair.
Romantic Caption : He walked into the living room and sat on the couch chair.

(b)

Figure 1. The comparison between factual caption and romantic caption. (**a**,**b**) show the factual caption and romantic caption for an image, respectively.

Based on the above discussion, we propose a novel stylized image captioning model, called ST-BR. We first use convolutional neural networks to extract image features, separate the image into several regions, and then use a style transformer (STrans) that integrates style information to process the extracted image information. The encoder of STrans models the relationship between image regions from high-level and low-level, respectively, which can effectively utilize the image's information. This method can better model the spatial relationship of image information and make the model definitely express factual image information. Besides, we introduce a style attention module (SAM) on the decoder part in STrans. Precisely, at each time step, SAM will calculate the current specified

style's attention weight when generating a word so that the generated captions will be more suitable for specific style factors. This method can effectively integrate multiple styles, and allowing the model to learn style knowledge from the corpus. It can improve the model's ability that controls the style of generated captions. Although STrans already can generate style captions, its capabilities still need to be further enhanced. Therefore, we use the back-reinforcement module to improve the model's ability to generate captions in an adversarial manner. We make the caption generator effectively integrate the relevance of the image and the closeness of the style via a back-reinforcement module. In this way, the generated captions not only reflects a specific writing style but also does not deviate from the factual image information. Specifically, in the pre-trained stage, the factual knowledge discriminator makes the caption generator strong and combines the image's factual content. In the fine-tuning stage, the style factor discriminator helps the caption generator closely integrate a specific style.

Our contributions in this paper are as follows:

1. We propose a stylized transformer structure. This structure efficiently matches style information through a multi-head style attention mechanism and generates stylized image caption rather than the factual caption.
2. We propose a back-reinforcement module to evaluate the degree of matching between the image and style information and the generated captions.
3. Based on 1 and 2, we design a stylized image captioning model based on adversarial learning. The game process enables the model to generate an accurate stylized caption based on learning the image's factual information.
4. We test our model on public datasets, and the results show the effectiveness of our proposed method.

2. Related Work

2.1. Traditional Image Captioning

The image captioning in deep learning starts with the CNN-LSTM that used convolutional networks (CNN) as an encoder and long short-term memory networks (LSTM) as a decoder [11,12] model. To improve performance, scholars introduce attention mechanisms, object detection, graph representation learning, etc. In 2017, Shetty et al. [13] and Dai et al. [14] proposed to train image captioning models via generative adversarial networks, respectively. They focused on the naturalness and diversity of generated captions. The former brings captions closer to human expression. They use adversarial training combined with an approximate Gumbel sampler to implicitly match the generated distribution to the human distribution, which improves the diversity of captions. The core idea of the latter is to adopt a conditional generative adversarial network (GAN). The general idea is consistent with Seq-GAN. The discriminator judges the generated captions and optimizes the model via policy gradient technology. In 2018, Anderson et al. [2] introduced an object detection algorithm and proposed a bottom-up and top-down attention mechanism. This method extracts the region of interest by Faster R-CNN [15], and calculate the bottom-up attention mechanism. They put the calculation steps of the bottom-up attention mechanism in the data processing part, simplifying the complexity of the model, and they implement the top-down attention mechanism by a two-layer LSTM structure. In the same year, Yao et al. [16] proposed a novel framework, which utilizes graph convolutional neural networks (GCNs) to model semantic and spatial information from an image. The method first builds graph structures based on the spatial and semantic relationships of the detected objects from images, then encode the regional features by the GCNs, and finally inputs it to the LSTM containing the attention mechanism. Li et al. 2019 [17] changed the transformer's internal structure, omitting the residual connection, layer normalization, and an embedding layer to adapt to the task of image captioning. In this way, the author bridges the gap between visual signals and semantic information. This work achieved the latest performance at the time. In 2020, Cornia et al. [18] proposed a mesh memory transformer for image captions. They replaced the connection between the encoder and the decoder in the original transformer with a mesh-like structure, which can better adapt to the image captioning's

multi-modal properties. This method shows the transformer can play an important role in image captioning.

2.2. Stylized Image Captioning

In 2017, Gan et al. [6] proposed an end-to-end training framework for automatically distilling styles in single-language text, and they also constructed a stylized caption data set. Researchers have improved the stylized image captioning framework's performance by introducing other fields' knowledge in subsequent work. In 2018, Mathews et al. [7] proposed a model for learning and generating visually relevant style captions from a large-scale style text corpus that has never aligned images. Their core idea is to separate semantics and styles. This work faces a single style(romantic). In the same year, Chen et al. [19] proposed a variant of LSTM. They adopted an adaptive learning method to solve the problem by comparing the model's predicted sub-probability distribution and the style model. Nezami et al. 2019 [20] proposed ATTEND-GAN, it includes an attention-based caption generator and adversarial training mechanism, which improved the stylistic flexibility of caption. They designed a novel training method that used the SentiCap to fine-tune the model. Guo et al. 2019 [9] proposed a multi-style image caption model based on the adversarial network, which mainly includes a style-dependent caption generator, a description discriminator, a style classifier, and a reverse translation module.

3. Method

3.1. Motivation

In the Internet age, various long or short videos flood our lives. Video understanding has become an essential task. A video is composed of several frames of images to regard image comprehension as a low-level video comprehension. When watching a video, we sometimes cannot obtain a high-level understanding of the video from the captions. Video captioning and image captioning technology can help us further understand their content. Compared with the factual caption, the stylized caption can better help us understand videos or images. The stylized captioning is for ordinary audiences, but the stylized caption technology can greatly help the visually impaired understand videos and images.

3.2. Overview

The goal of stylized image captioning is to generate captions expressed in a specific style while effectively using the image's factual information. This paper integrates style information into the transformer [10] and encapsulates it into a new transformer structure. We apply the style-transformer to the stylized image captioning for the first time, and the adversarial training method further optimizes the model. Although the encoder and decoder of style-transformer are slightly different, the overall structure is symmetrical. Figure 2 is an overview of our model. Next, we will elaborate on the style transformer module, back-reinforcement module, and other important information.

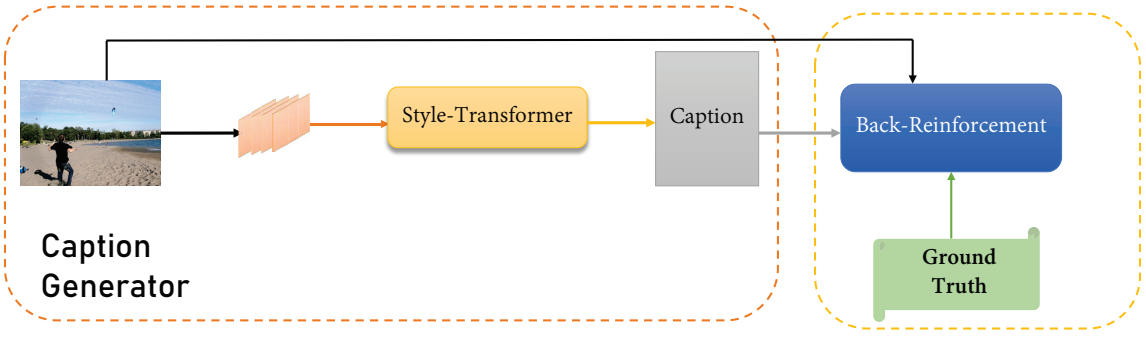

Figure 2. The overview of ST-BR, ST-BR includes the caption generator and back-reinforcement module.

3.3. STrans: A Better Framework to Understand Images

3.3.1. Encode Images: Effectively Utilize the Information of Images

In this paper, the model extracts feature information in the image via convolutional neural networks (e.g., ResNet152 [21], VGGNet [22]) and obtains a set of image features: $\mathbf{R} = \{r_1, r_2, \ldots, r_n\}, r_i \in \mathbb{R}^D$. The STrans encoder further encodes the image information hierarchically, making the model efficiently use its information. Before input into the STrans encoder, the model converts the image features into three vectors of query, key, and value, obtained through linear mapping. The calculation of visual self-attention can be expressed as follows (the following calculation takes a single coding layer as an example):

$$Z = Attention(W^Q R, W^K R, W^V R) \tag{1}$$

In order to obtain different information from multiple angles, we still need to calculate multi-head attention:

$$M^l = Concat(Z_1, \cdots, Z_h) \tag{2}$$

where h is the number of heads, M^l represents the multi-head attention of the first layer encoder and l is the number of layers, and Concat() denotes concatenation operation. Next, the model feeds it into layer normalization, defined as follows:

$$FFN = U\sigma(AM^l + b) + c \tag{3}$$

In the Equation (3), σ denotes the Sigmoid function, b and c is the bias matrix, U and A is the learnable parameter. After that, the module feeds the obtained results, and the original image features into layer standardization operations through residual connection, which is defined as follows:

$$R_1 = AddNorm(M^l) \tag{4}$$

$$R_2 = AddNorm(FFN(R_1)) \tag{5}$$

Our encoder base on the above structure, stacked in multiple layers in sequence. Therefore, the input of layer $t$is the output of layer $t-1$. We understand the image information from a higher level by stacking the layer structure and continue using and refining the relationship between the image regions encoded in the previous layer. The superimposed layer's encoder can get an output from each layer, defined as: $\widetilde{R} = (R_2^1, \ldots, R_2^N)$.

3.3.2. Style Attention Module: Closely Integrate Style Factors

The style attention module is based on the decoder's multi-head attention in STrans and introduces a multi-style module for calculating style attention. The model utilizes a one-hot vector to express a multi-style module. The module (in Figure 3) uses m + 1-dimensional vectors to represent m different styles. In the existing dataset, captions' styles are mainly romantic, humorous (FlickrStyle10K), etc. The decoding layer and the STrans coding layer have a similar structure, but the differences are style attention module and encoder-decoder attention module. We input four parts into the decoding layer, including the words generated at the previous time step, encoded image information, style information, and position encoding. The model inputs the word generated at the previous time step, style information, and position code to calculate style attention. Among them, position-coding alleviates the long-term dependence in sequence tasks. The overall calculation process is as follows (take a single decoding layer as an example) :

$$SA^l = Concat(Multihead(W^{DQ}(x_{t-1}, S_i), W^{DK}(x_{t-1}, S_i), W^{DV}(x_{t-1}, S_i))W^d) \tag{6}$$

$$Z_D = AddNorm(SA^l) \tag{7}$$

where SA^l denotes style attention of the i-th, S_i is the i-th style.

$$EDA^l = Concat(Attention(W^{Q'} Z_D, W^{K'} \tilde{R}, W^{V'} \tilde{R}) W^{ed}) \tag{8}$$

$$Z_D^{'} = AddNorm(EDA^l) \tag{9}$$

$$Z_D^{''} = AddNorm(FFN(Z_D^{'})) \tag{10}$$

EDA^l represents encoder-decoder attention. Next, the model feeds the normalization layer's result into the Softmax layer to calculate the probability distribution. The model performs a table lookup operation according to the output probability distribution, outputs the word at the current moment, and combines the output at t moments into a sentence.

$$p = Soft\max(w_p Z_D^{''} + b_p) \tag{11}$$

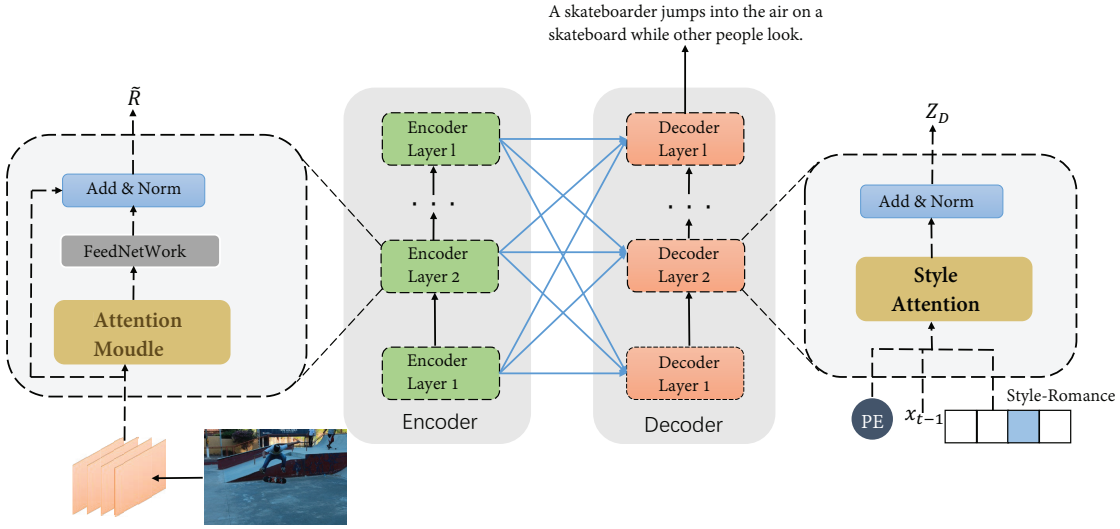

Figure 3. The inner structure of style-transformer, the left-hand side shows the process of image encoding and the right-hand side is style attention module.

3.4. Back-Reinforcement: Further Optimize the Caption Generation Module

In the above work, we fully described the generation process of stylized image captions. Obviously, for what we want to achieve, it is not enough to rely on previous work. Our preliminary work can only make the model generate captions with certain style factors. However, we cannot measure the pros and cons of generating captions. Whether the captions generated by this method are consistent with the image's factual information and how much difference exists between them. If the model generates captions consistent with a specific style and deviates from the image information, such captions are of no value. The opposite is also true. Based on these considerations, we used the idea of adversarial learning to design a back-reinforcement module (in Figure 4).

The essence of the back-reinforcement module is a discriminator. The style-transformer and the back-reinforcement module make up the framework of a generative adversarial network (GAN). The core idea of GAN is the game. The generator aims to generate captions that can confuse the discriminator. The goal of the discriminator is that accurately discriminate the generated fake captions. In their mutual game, the caption generation model can produce better quality captions. The module includes factual knowledge discriminator and style factor discriminator, which will be introduced in detail next.

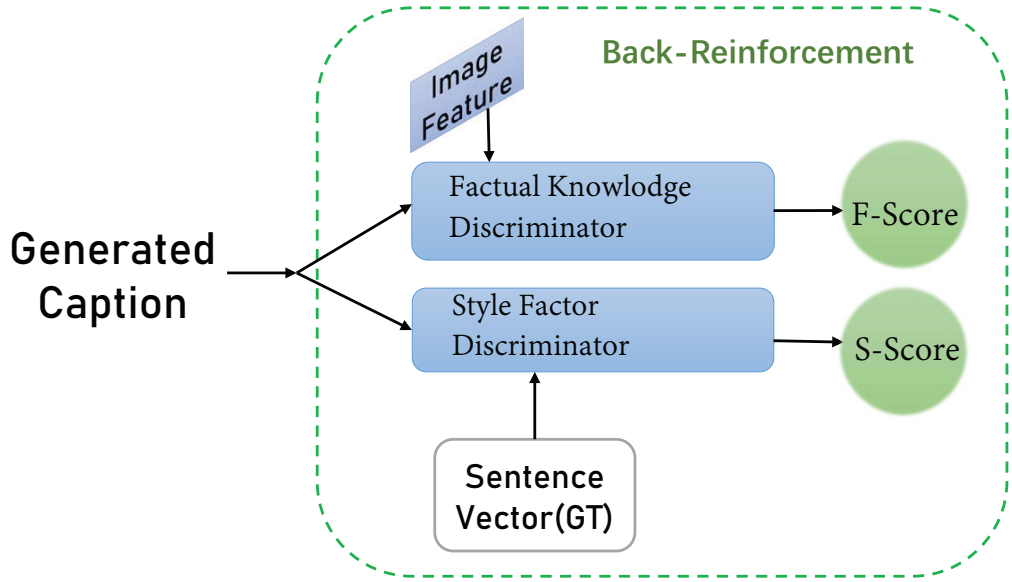

Figure 4. The overview of the back-reinforcement module, sentence vector denotes manually annotated caption from human (ground truth).

3.4.1. Factual Knowledge Discriminator (FKD): Let the Model Make More Use of Image Information

The generated captions should express the image's factual information, which is very important in image captioning. We designed a discriminator based on the image's factual information, called FKD (in Figure 5). Image captioning is a cross-modal technology. Image information and generated caption information are two modalities. Therefore, in order for the discriminator to effectively play its role, we must consider this issue when designing its internal structure. What kind of structure should we adopt to compare the consistency of the two kinds of information. Finally, we choose the convolutional neural networks to process image and caption and then input its result into a fully connected layer with the sigmoid. We can get a predicted score to measure the degree of consistency between the image's real information and the generated captions. The input of FKD includes original image features and generated captions. In order for CNN to effectively obtain the feature information of both of them, we first need to map image features and sentence features to the same space and construct a new feature map, expressed as:

$$\Gamma = \bar{r} \oplus Ex_2 \oplus Ex_2 \oplus \cdots \oplus Ex_L \tag{12}$$

Among them, $\bar{r} = CNN(I)$ is the image feature, L is the length of the generated captions, and E is the embedding matrix. Its function is to align the dimensions of the two and $\oplus$ is the connection operation. The model sets filters of different sizes in each channel.

This method can capture different information levels in the new feature map, which can better measure the relevancy between the image and the generated captions. Through this method, the model can further strengthen the effect of image information.

Take one of the channels as an example, and the window size is $D \times L$, we first use it to capture the features of Γ and built a new feature map, expressed as: $f = [f_1, f_2, \ldots, f_{(T-l+2)}]$. The calculation method of each element in is $f_i = ReLU(\kappa * \Gamma_{(i:i+l-1)} + b)$, where κ represents the channel, * represents the convolution operation, b is the bias weight, and $ReLU()$ is the rectified linear unit. After that, the module further extracts the information on the feature map by using the maximum pooling. Next, we connect the pooled results of different channels into a feature vector F. In order to improve the performance of the model, inspired by [23] before inputting the multilayer perceptron, we need to process F as follows:

$$\delta = \sigma(W_T \cdot F + b_T) \tag{13}$$

$$H = \mathrm{ReLU}(W_H \cdot F + b_H) \tag{14}$$

$$F^{'} = \delta \odot H + (1 - \delta) \odot F \tag{15}$$

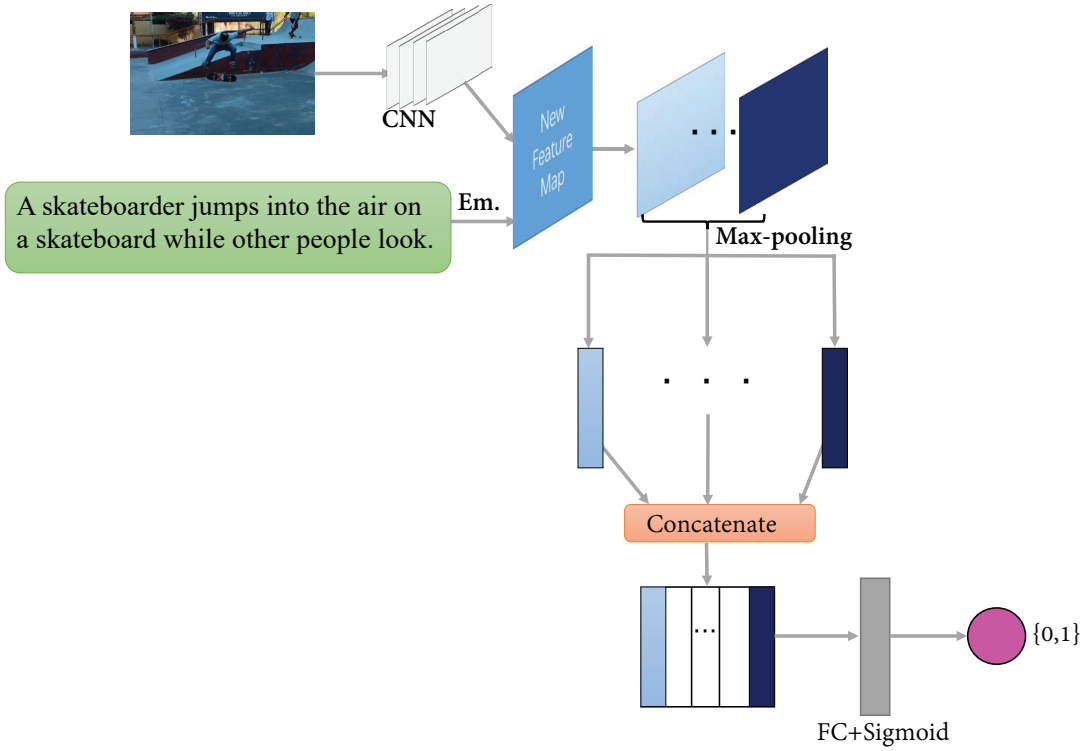

Figure 5. The factual knowledge discriminator, Em. is the embedding layer.

3.4.2. Style Factor Discriminator (SFD): Make the Model Closer to a Specific Style

In addition to using factual knowledge of images, the generated captions also need to be highly compatible with specific style factors. FKD (in Figure 6) solves the problem of relevancy between the input image and the generated captions. Although we added a style factor to STrans, the model still needs to further enhance the captions' style. Therefore, we propose a style factor discriminator, called SFD. The module's core idea is to compare the generated captions with the real labeled captions to obtain an evaluation score. Unlike FKD, because there is a kind of modality (text), LSTM is used as the main structure of SFD, and then it is connected to the fully connected layer with the sigmoid function.

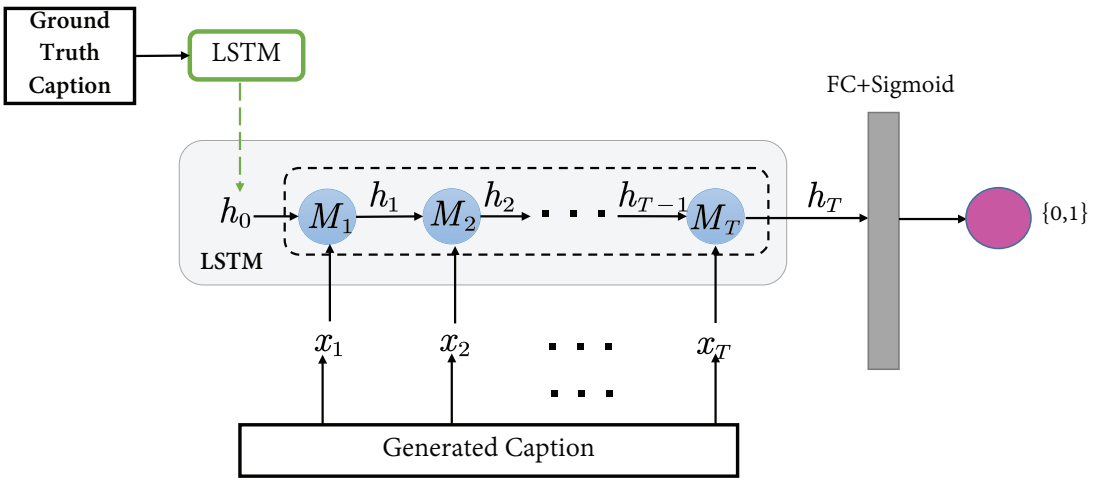

Figure 6. The style factor discriminator.

Specifically, the module encodes the truly annotated sentence by encoding LSTM (Enc-LSTM) and uses the hidden state at the final moment to represent the sentence (S_{real}). At the initial time step 0, we use S_{real} to initialize the hidden state of LSTM (Dis-LSTM). Both Enc-LSTM and Dis-LSTM are standard LSTM structures. After that, the model inputs the generated captions into Dis-LSTM. Formulated as follows:

$$h_{t+1} = \begin{cases} \text{Dis} - \text{LSTM}\left(s_{\text{real}}, h_t\right) & t = 0 \\ \text{Dis} - \text{LSTM}\left(Ex_t, h_t\right) & t = (1, 2, \ldots, T) \end{cases} \tag{16}$$

The hidden state h_{T+1} at the final moment is input to the fully connected layer with the Sigmoid function. Finally, we can get a score that measures similarity:

$$S_s = \sigma(W_s \cdot h_{T+1} + b_s) \tag{17}$$

3.5. Training Strategy

We train the model in a two-task learning manner due to the lack of large-scale paired image-stylized captioning data sets. We divide the training into two stages: the pre-training stage and the fine-tuning stage. The pre-training stage's main task is to enable the model to recognize and efficiently use the image's factual information, which is to keep the generated captions consistent with the factual information of the image. The model uses traditional data sets for training at this stage, for example, MS COCO, Flickr30k, etc. The fine-tuning stage aims to make the captions generated by the model closely follow specific style factors. At this stage, the model is trained by stylized datasets, for example, FlickrStyle10K.

In the two training phases, we train in a confrontational manner. Besides, we calculate the loss function of each stage separately and minimize the loss. In the pre-training stage, loss needs to calculate two parts: the caption generator and the factual knowledge discriminator in the back-reinforcement module, defined as the following formula:

$$Loss_{pre} = -\frac{1}{2}E_{(R,S_{rs})\in S_{Tdata}} \log(D(R, S_{rs})) - \frac{1}{2}E_{(R,S_{fs})\in S_{Gdata}}[log(1 - D(G(R, S_{fs})))] \tag{18}$$

In Equation (18), S_{Tdata} is a real caption data set, S_{rs} represents a real caption data set, S_{Gdata} is a generated caption set, and S_{rs} represents a generated pseudo caption data. Similarly, the fine-tuning stage also needs to calculate the loss of two parts, defined as follows:

$$\text{Loss}_{fin} = -\frac{1}{2}E_{x\sim S_x} \log D(x) - \frac{1}{2}E_{gs\sim S'_g} \left[\log\left(1 - D\left(G\left(g_s\right)\right)\right)\right] \tag{19}$$

S_{st} is the real style caption data set, S'_{gs} is the generated style caption set, and g_s represent the generated captions. We adopted the standard GAN [24] and Self-Critical [25] training methods throughout the training process.

4. Experiment

This section will detail the experimental part, including previous work, implementation details, result analysis, and ablation studies.

4.1. Preparatory Work

4.1.1. Datasets

This paper aimed to achieve our experiment on two datasets: MSCOCO 2014 [26] and FlickrStyle10K [6]. **MS COCO 2014** is a large-scale vision task dataset, which can be used for tasks such as object detection, segmentation, human keypoint detection, and image captioning. In terms of image captioning, the dataset had 164,062 images, and each image corresponded to five captions.

Among them, 82,783, 40,504, and 40,775 images were used for training, verification, and testing. To compare with the results obtained in other work, we adopted the method of dividing the data set by Karpathy et al. [11]. Their method only included the original training set and validation set, a total of 123,287 images. The method included 113,287 images for training, 5000 for verification, and 5000 for testing.

FlickrStyle10K contained 10,000 Flickr images and an unpaired text corpus, including two texts: romantic and humorous. The author only released the 7 K training set; therefore, we randomly selected 6000 images for training and the rest for testing. We used Hum and Rom to represent Humorous and Romantic styles, respectively.

We used MSCOCO 2014 to pre-train the model, and then fine-tuning stage train the model by FlickrStyle10K.

4.1.2. Evaluation Metrics

We evaluated our model for stylized image captioning in two aspects: the relevancy of the stylized caption with the image and the caption's language style. Regarding relevance, we adopted traditional evaluation methods of image captioning, which mainly includes BLUE-n [27] (B-1&B-3), CIDEr [28], and METEOR [29]. Besides, for language style, we started with two evaluation points, fluency and style accuracy. The details of the two were as follows:

- **Fluency** For evaluating sentence fluency, we utilized a language modeling tool (SRILM [30]) to achieve it. SRILM calculates the perplexity generated sentences using the trigram language model trained on the respective corpus [9]. SRILM calculates the perplexity scores (ppl.) of each style generated caption. The smaller the value of the SRILM score denotes, the lower the perplexity of captions generated by the model, the more fluent the sentence, and the better the model's performance.
- **StyleAccuracy** Style classification accuracy (cls.) is the proportion of captions that accurately integrate style factors in all captions. We calculate style accuracy by a style classifier, which is pre-trained on FlickrStyle10K and MSCOCO 2014 datasets.

4.1.3. Compared Models

For proving the effectiveness of our model, we selected some models for comparison. There is mainly the following baseline:

- **NIC-TF**: This method uses each style of data set to fine-tune on the NIC model [11].
- **StyleNet** [6]: The work proposes an end-to-end learning framework that can automatically distill the style factors in a single language text.
- **SF-LSTM** [19]: The work proposes a variant of LSTM; its function is to obtain factual and style knowledge. In addition, during the training process, they proposed to use the actual output of the parameters as a guide. It is a supervised method.
- **MSCap** [9]: This paper trains a model to implement the task of learning multi-style captions from unpaired data.

Compared with these methods, our model had an evident difference with them. NIC-TF was fine-tuned on the style corpus, and there was no specific module to learn style knowledge. In our model, the stylized attention mechanism solved this problem. StyleNet learned the styles in the corpus by knowledge distillation, it learned through the weight update method. The model could only learn one style at a training process, but our method could learn multiple styles at the same one. SF-LSTM is a supervised learning method that utilizes two sets of matrices to capture factual information and style information. This method relies heavily on paired data. MSCap uses the traditional convolutional neural network to encode image information, which cannot effectively model the factual image information. We encoded the high-level and low-level information of the image

via style-transformer, which made our method correctly express the image information. Besides, our method strengthened the model from two aspects: factual information and style information.

4.2. Implementation Details

We implemented our code on the Pytorch platform, one of the most popular deep learning frameworks. Next, we will introduce some parameter settings in the experiment. In the data processing, we extracted image features by the pre-trained ResNet152 and obtain 7*7*2048 feature maps. We set the dimension of each layer of STrans to 512, the number of headers to 8, and stack 6 layers of network structure. In two training stages, we used the Adam [31] optimization algorithm to optimize the model, and set the weight decay. The learning rate was 5×10^{-4}. The batch size of the pre-training stage and the fine-tuning stage were set to 64 and 96, respectively, and the word embedding used 300-dimensional GloVe [32]. We used the beam search in the word generation step, and its size was set to 5. Dropout rate is 0.5. We regarded factual as a style; there were three styles: factual (no evaluation), romantic, and humorous. For verification and testing, we set the size of the beam search to 5. In the back-reinforcement module, the size of the LSTM hidden state in SFD was set to 512. Our model worked on two 16 GB Tesla v100-PICE GPUs for 89 h, and the training process lasted for 20 epochs.

4.3. Result Analysis

4.3.1. Quantitative Analysis

We evaluated the stylized captions from three aspects: the degree of relevance between the stylized captions and the image, the fluency of the stylized captions, and the accuracy of the caption style. The latter two focus on evaluating the caption language style, and that is the essential evaluation index in stylized image captioning. To better compare, we set up two different experimental methods of a single style and multiple styles.

The relevance of captions and images: We adopted the evaluation metrics of the image captioning to evaluate the correlation between style captions and images. It is worth noting that, because the traditional evaluation method uses the N-Grams method to calculate the score, and the stylized vocabulary in the stylized captions is unique, the traditional way was not very good for the accurate assessment of stylized captioning. Table 1 shows the performance of the single style method. Since we used a weakly supervised method, the experimental results were worse than the supervised method (SF-LSTM) in the traditional evaluation method. However, compared with StyleNet, which also uses weak supervision methods, our model performed better in all traditional indicators. It also shows that our method could better model image information and express it reasonably. In terms of multiple styles, we can see from Table 2 that our performance improved compared to MSCap.

Linguistic Style of Captions We evaluated the style captions' consistency with the specified style by calculating the perplexity score and the caption style's accuracy score. The smaller the value of the confusion score, the smoother the generated captions, which means that the captions were more closely related to the style. In Tables 1 and 2, whether it was a single-style model or a multi-style model, our model achieved better results compared with the baseline. First, our model had a lower perplexity score, which means that our model could generate smoother captions. Besides, our model also achieved good results in terms of style classification accuracy. It shows that our method could better integrate style factors and effectively guide the generation of stylized captions.

Table 1. Performance of single style method. Rom represents a romantic style, and Hum denotes a humorous style. The experimental effect is obtained through an unpaired corpus. B@n, M, C, ppl, and cls represent Blue-1, Blue-3, METEOR, CIDEr, perplexity, and style classification accuracy (%), respectively. In addition, the smaller the ppl score, the generated captions are more fluent.

Model	Style	B-1	B-3	M	C	ppl()	cls
NIC-TF	Rom	26.9	7.5	11.0	35.4	27.7	82.6
	Hum	26.3	7.4	10.2	35.1	31.8	80.1
StyleNet	Rom	13.3	1.5	4.5	7.2	52.9	37.8
	Hum	13.4	0.9	4.3	11.3	48.1	41.9
SF-LSTM	Rom	27.8	8.2	11.2	37.5	-	-
	Hum	27.4	8.5	11.0	39.5	-	-
ST-BR	Rom	22.5	3.7	6.6	21.7	17.3	95.2
	Hum	20.4	4.3	7.5	18.2	16.1	96.1

Table 2. Performance of multi-style method.

Model	Style	B-1	B-3	M	C	ppl	cls
MSCap	Rom	17.0	2.0	5.4	10.1	20.4	88.7
	Hum	16.3	1.9	5.3	15.2	22.7	91.3
ST-BR	Rom	18.1	2.3	5.1	11.5	18.8	90.1
	Hum	17.6	2.1	5.5	16.7	18.1	92.5

4.3.2. Qualitative Analysis

To demonstrate the effectiveness of the model, we list an example of model generation in Figure 7. We believe that romance is a subjective feeling, and humor is an emotion that makes others happy. In Figure 7, we found that the captions generated by ST-BR not only effectively expressed image information, but also described subjective feelings from a human perspective, such as the phrases "get where she is going" and "find his families". From a humorous point of view, we could feel happy emotions from the phrases "joyfully skipping right into the water" and "looking for clowns".

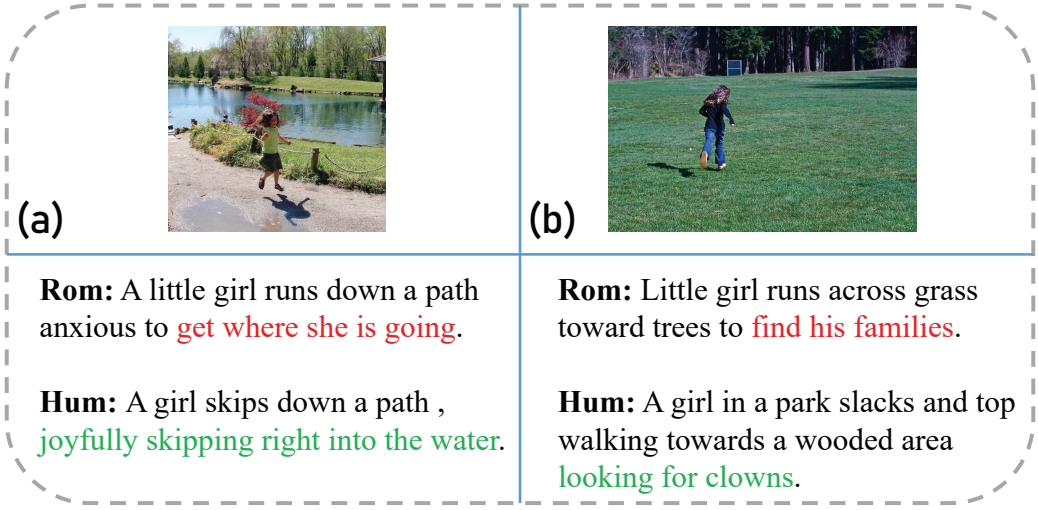

Figure 7. (**a**,**b**) Show the stylized captions generated by ST-BR for an image, respectively, the captions include a romantic caption and humorous one. Colored fonts reflect the language style of captions.

4.3.3. Ablation Studies

Our model consisted of a caption generator (STrans) and a reverse enhancement module (BR). The factual knowledge discriminator (FKD) and the style factor discriminator (SFD) constitute the back-reinforcement module. Therefore, we needed to verify their effects on the entire model. The structure of all our variant models is shown in Table 3. We divided the model into the following parts (analyzed the experimental results of a single model):

- **STrans**: To verify the multi-discriminator module's function, we removed the reverse enhancement module and only used the caption generator module.
- **ST-FKD**: We removed SFD to verify the style factor discriminator's guiding role on style.
- **ST-SFD**: To verify the effect of the factual knowledge discriminator, we only kept the style factor discriminator in the reverse enhancement module.

From the results in Table 3, we can see that the performance of STrans was inferior to ST-FKD in terms of the correlation between captions and images, but inferior to ST-SFD in terms of caption language style. Besides, although the captions generated by ST-FKD had high consistency with the image, its performance in terms of style was inferior. In this regard, ST-SFD had better performance, but the captions it generated were far from the image's factual information. In the end, ST-BR with a full model had the best performance.

Table 3. The results of ablation studies.

Method	Style	B1	B3	M	C	ppl	cls
STrans	Rom	12.2	1.3	4.7	8.0	54.3	35.6
	Hum	13.0	0.7	4.3	13.4	46.2	43.5
ST-FKD	Rom	18.9	2.5	5.5	16.7	59.3	30.4
	Hum	17.0	1.6	6.0	15.9	55.2	29.9
ST-SFD	Rom	10.0	0.8	3.2	5.7	33.2	76.9
	Hum	9.8	0.5	3.0	6.0	30.1	88.0
ST-BR	Rom	22.5	3.7	6.6	21.7	17.3	95.2
	Hum	20.4	4.3	7.5	18.2	16.1	96.1

5. Conclusions

In this paper, we have integrated multi-style information and encapsulated the transformer structure and multi-style modules as a style transformer. Besides, we combined the approach of adversarial learning to design a novel stylized image captioning framework, called ST-BR. ST-BR is composed of a caption generation module and a back-reinforcement module. Among them, the factual knowledge discriminator and the style factor discriminator constitute the back-reinforcement module. In the caption generator module, ST-BR learns style factors through the style attention module. In the back-reinforcement module, the factual knowledge discriminator calculates the relevancy score between the generated captions and the image. The generated captions can better express the factual information of the image. The style factor discriminator is used to discriminate the language style of the generated captions. Through the analysis of experimental results, we confirmed that the captions generated by our model could not only be consistent with the factual information of the image but also have a specific style. A large number of experiments show the effectiveness of our model.

Nowadays, the short video is entering everyone's life, bringing new opportunities and challenges to caption technology. We can regard Image captioning as part of video captions. In the future, we can transfer our work to stylized video captioning. It will further promote people to understand videos.

Besides, for our method, we believe that further research can be done from the following aspects. First, we can construct a paired dataset, which includes images and their corresponding stylized

captions. This work will greatly promote the development of stylized image captioning. The second point is that we conduct research on the application of this algorithm in the industrial field.

Author Contributions: Data curation, Q.L. and G.L.; formal analysis, Z.Y. and Q.L.; funding acquisition, Z.Y.; investigation, G.L.; methodology, Z.Y. and Q.L.; software, Q.L.; supervision, Z.Y. and G.L.; writing—original draft, Q.L.; writing—review and editing, Z.Y. and G.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by National Key R&D Program of China [No. 2019YFB1404700].

Acknowledgments: This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFB1404700.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chen, S.; Jin, Q.; Wang, P.; Wu, Q. Say As You Wish: Fine-Grained Control of Image Caption Generation With Abstract Scene Graphs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 9962–9971.
2. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 6077–6086.
3. Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; Chua, T.-S. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 26 July 2017; pp. 5659–5667.
4. Pan, Y.; Yao, T.; Li, Y.; Mei, T. X-Linear Attention Networks for Image Captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–18 June 2020; pp. 10971–10980.
5. Jing, B.; Xie, P.; Xing, P.E. On the Automatic Generation of Medical Imaging Reports. *arXiv* **2018**, arXiv:1711.08195.
6. Gan, C.; Gan, Z.; He, X.; Gao, J.; Deng, L. StyleNet: Generating Attractive Visual Captions With Styles. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3137–3146.
7. Mathews, A.; Xie, L.; He, X. SemStyle: Learning to Generate Stylised Image Captions Using Unaligned Text. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 8591–8600.
8. Zhang, W.; Ying, Y.; Lu, P.; Zha, H. Learning Long- and Short-Term User Literal-Preference with Multimodal Hierarchical Transformer Network for Personalized Image Caption. In Proceedings of the AAAI, New York, NY, USA, 7–12 February 2020. Available online: https://doi.org/10.1609/aaai.v34i05.6503 (accessed on 30 November 2020).
9. Guo, L.; Liu, J.; Yao, P.; Li, J.; Lu, H. MSCap: Multi-Style Image Captioning With Unpaired Stylized Text. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–21 June 2019; pp. 4204–4213.
10. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS), Long Beach, CA, USA, 4–10 December 2017; pp. 5998–6008.
11. Karpathy, A.; Li, F.-F. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; Volume 2015, pp. 3128–3137.
12. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3156–3164.
13. Shetty, R.; Rohrbach, M.; Hendricks, L.A.; Fritz, M.; Schiele, B. Speaking the Same Language: Matching Machine to Human Captions by Adversarial Training. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.

14. Dai, B.; Fidler, S.; Urtasun, R.; Lin, D. Towards diverse and natural image descriptions via a conditional gan. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2989–2998.
15. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*; IEEE Computer Society: Los Alamitos, CA, USA, 2017; Volume 39, pp. 1137–1149. [CrossRef]
16. Yao, T.; Pan, Y.; Li, Y.; Mei, T. Exploring visual relationship for image captioning. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 684–699.
17. Li, G.; Zhu, L.; Liu, P.; Yang, Y. Entangled Transformer for Image Captioning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 8928–8937.
18. Cornia, M.; Stefanini, M.; Baraldi, L.; Cucchiara, R. Meshed-Memory Transformer for Image Captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–18 June 2020; pp. 10578–10587.
19. Chen, T.; Zhang, Z.; You, Q.; Fang, C.; Wang, Z.; Jin, H.; Luo, J. "Factual" or "Emotional": Stylized Image Captioning with Adaptive Learning and Attention. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 519–535.
20. Nezami, O.M.; Dras, M.; Wan, S.; Paris, C.; Hamey, L. Towards Generating Stylized Image Captions via Adversarial Training. In *Pacific Rim International Conference on Artificial Intelligence*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 270–284.
21. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
22. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
23. Yu, L.; Zhang, W.; Wang, J.; Yu, Y. SeqGAN: Sequence generative adversarial nets with policy gradient. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17), San Francisco, CA, USA, 4–9 February 2017; pp. 2852–2858.
24. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems 27 (NIPS), Montreal, QC, Canada, 13 December 2014; pp. 2672–2680.
25. Rennie, S.J.; Marcheret, E.; Mroueh, Y.; Ross, J.; Goel, V. Self-Critical Sequence Training for Image Captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 7008–7024.
26. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Doll'ar, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
27. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
28. Vedantam, R.; Zitnick, C.L.; Parikh, D. Cider: Consensusbased image description evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4566–4575.
29. Banerjee, S.; Lavie, A. Meteor: An automaticmetric for mt evaluation with improved correlation with human judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, MI, USA, 29 June 2005 ; pp. 65–72.
30. Stolcke, A. Srilm-an extensible language modeling toolkit. In Proceedings of the Seventh International Conference on Spoken Language Processing, Denver, CO, USA, 16–20 September 2002. Available online: https://isca-speech.org/archive/archive_papers/icslp_2002/i02_0901.pdf (accessed on 30 November 2020).

31. Kingma, D.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
32. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; Volume 14, pp. 1532–1543.