"Factual" or "Emotional": Stylized Image Captioning with Adaptive Learning and Attention

Tianlang Chen[1][0000−0002−6355−6474], Zhongping Zhang[1], Quanzeng You[3], Chen Fang[2], Zhaowen Wang[2], Hailin Jin[2], and Jiebo Luo[1]

[1] University of Rochester, {tchen45,jluo}@cs.rochester.edu, {zzhang76}@ur.rochester.edu
[2] Adobe Research, {cfang,zhawang,hljin}@adobe.com
[3] Microsoft Research, {quyou}@microsoft.com

Abstract. Generating stylized captions for an image is an emerging topic in image captioning. Given an image as input, it requires the system to generate a caption that has a specific style (e.g., humorous, romantic, positive, and negative) while describing the image content semantically accurately. In this paper, we propose a novel stylized image captioning model that effectively takes both requirements into consideration. To this end, we first devise a new variant of LSTM, named style-factual LSTM, as the building block of our model. It uses two groups of matrices to capture the factual and stylized knowledge, respectively, and automatically learns the word-level weights of the two groups based on previous context. In addition, when we train the model to capture stylized elements, we propose an adaptive learning approach based on a reference factual model, it 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. We evaluate our model on two stylized image captioning datasets, which contain humorous/romantic captions and positive/negative captions, respectively. Experiments shows that our proposed model outperforms the state-of-the-art approaches, without using extra ground truth supervision.

Keywords: stylized image captioning, adaptive learning, attention model

1 Introduction

Automatically generating coherent captions for images has attracted remarkable attention for its strong applicability, such as picture auto-commenting [23] and helping blind people to see [11]. This task is often referred to as image captioning, which combines computer vision, natural language processing and artificial intelligence. Most recent image captioning systems focus on generating an objective, neutral and indicative caption without any style characteristics, which is defined as a factual caption. However, the art of language motivates researchers to generate captions with different styles, which can give people different feelings

when focusing on a specific image. The "style" can refer to multiple meanings. For example, as shown in Figure 1, in terms of the fashion of the caption, caption style can be either "romantic" or "humorous". In addition, in terms of the sentiment it brings to people, caption style can be either "positive" or "negative". Without doubt, generating such kinds of captions with different styles will greatly enrich the expressibility of the captions and make them more attractive.

Ideally, a high-performing stylized image captioning model should satisfy two requirements: 1) it generates appropriate stylized words/phrases in appropriate positions of the caption, 2) it still describes the image content accurately. Focused on stylized caption generation, existing state-of-the-art work [28][9] train their captioning models based on two datasets separately, a large dataset with paired images and ground truth factual captions, and a small dataset with paired images and stylized ground truth captions. From the large factual dataset, the model is learned to generate factual captions that can correctly describe the images; from the small stylized dataset, the model is learned to transform factual captions to stylized captions by incorporating suitable non-factual words/phrases at correct positions of the caption. In the training and predicting process, how to effectively take these two aspects into consideration is paramount for the model to generate high quality stylized captions.

To combine and preserve the knowledges learned from both factual and stylized dataset, Gan et al. [9] propose a factored LSTM, which factorizes matrix $W_{x\cdot}$ into three matrices ($U_{x\cdot}$, $S_{x\cdot}$, $V_{x\cdot}$). $U_{x\cdot}$ and $V_{x\cdot}$ are updated by the ground truth factual captions while $S_{x\cdot}$ is updated by ground truth captions with a specific style. In the predicting process, $U_{x\cdot}$, $S_{x\cdot}$ and $V_{x\cdot}$ are combined to generate the stylized caption. Since $U_{x\cdot}$ and $V_{x\cdot}$ preserve the factual information and $S_{x\cdot}$ preserves the stylized information, the model can thus generate stylized captions that correspond to input images. However, for both the training and predicting processes, factored LSTM cannot differentiate whether paying more attention to the fact-related part (i.e. $U_{x\cdot}$ and $V_{x\cdot}$) or the style-related part (i.e. $S_{x\cdot}$). It is natural that when the model focuses on predicting a stylized word, it should pay more attention to the style-related part, and vice versa. Mathews et al. [28] consider this problem and propose Senticap, which consists of two parallel LSTMs – one updated by the factual captions and one updated by the sentimental captions. When predicting a word, Senticap obtains the result by weighting the predicted word probability distributions of the two LSTMs. However, directly weighting the high level probability distributions can be too "coarse" in that it doesn't consider the low level attention effect on stylized and factual elements. In addition, Senticap obtains the weights of the two distributions by predicting the sentiment strength of the current word. In this step, it uses the extra ground truth word sentiment strength label, which is unavailable for other datasets.

In this paper, we propose a novel stylized image captioning model. In particular, we first design a new style-factual LSTM as a core building block of our model. Compared with factored LSTM, it combines fact-related and style-related parts of LSTM in a different way and incorporates self-attention for this two parts. More concretely, for both input word embedding feature and input

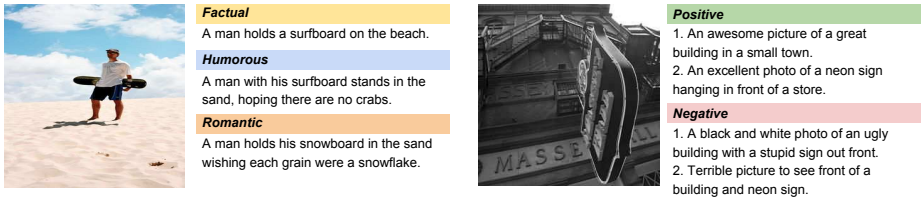

Fig. 1. Examples of stylized image captions. Besides factual captions, there can be four kinds of stylized captions that correspond to humorous, romantic, positive and negative styles, respectively.

hidden state of LSTM, we assign two independent groups of matrices to capture the factual and stylized knowledges, respectively. At each time step, it feeds an effective attention mechanism to weight the importance of the two groups of parameters based on previous context information, and combines the two groups of parameters by weighted-sum operation. In addition, to help the model preserve factual information while learning from stylized captions, we develop an adaptive learning approach that feeds a reference factual model as a guidance. At each time step, the model can adaptively learn whether to focus more on the ground truth stylized label or on the factual guidance, based on the similarity between the outputs of the real stylized captioning model and the reference factual model. Overall, both improvements help the model capture and combine the factual and stylized knowledge in a better way.

In summary, the main contributions of this paper are:

- We propose a new stylized image captioning model, with a core building block named style-factual LSTM. Style-factual LSTM incorporates two groups of parameters with dynamic attention weights into an LSTM, to adaptively adjust the relative attention weights between the fact and style-related parts.
- We develop a new learning approach to training the model on stylized captions, which adds the factual output of reference model as a guidance. The model can automatically adjust the strength of guidance based on ground truth stylized caption and reference model output without using additional information.
- Our model outperforms the state-of the-art methods on both image style captioning and image sentiment captioning task, in terms of both the relevance to the image and the appropriateness of the style.
- We visualize the corresponding attention weights for both the style-factual LSTM and the adaptive learning approach, and show explainable improvements in the results.

2 Related Work

Stylized image captioning is mainly related to two research topics: image captioning and style transfer. In this section, we provide the background of image captioning, attention model and style transfer, respectively.

Image Captioning. Image captioning has received much attention in recent years due to the advances in computer vision and natural language processing. Early image captioning methods [8][7][18][22][19][21][6] generate sentences by combining words which are extracted from corresponding images. A downside of these methods is that their performance is limited by empirical language models. To alleviate the problem, retrieval-based frameworks are developed [20][31][14][19]. They first retrieve similarity images of the input image from a database, then generate new descriptions for the query image by using the captions of retrieved images. However, this kind of approach relies heavily on the image database. Modern approaches [17][5][26][4][27][42][44][40] consider image captioning as a machine translation problem. Vinyals et al. [42] propose an encoder-decoder framework. Many improved approaches [17][5][26][29][44][40] are developed based on this encoder-decoder framework. The differences between these methods often lie in the architecture of recurrent neural network.

Attention Model. Recent successes of attention models [38][33][13][32][2] motivate many researchers to apply visual or language attention models [44][29] [24][37][45][1] to the image captioning task. Top-down visual attention models are first widely used [29][43][44][39]. The attention models enable deeper image understanding by assigning different attention weights to different image regions. Bottom-up and top-down combined attention models [45][1] are also proposed to take even one step further. In [24], the authors propose a novel adaptive attention model with a visual sentinel. This model not only can determine where to attend to in images, but also adaptively decide whether it needs to attend the image or to the LSTM decoder according to different words. Motivated by this work, we develop a novel joint style-factual attention architecture to make the model adaptively learns from the factual part and stylized part.

Style Transfer. Most style transfer works [10][16][30][41] focus on image style transfer. These works utilize the Gram matrix of hidden layers to measure the distance between different styles. In the meantime, pure text style transfer is making breakthrough as the development of nature language processing. For example, Shen et al. [35] propose a cross-alignment method to transfer text into different styles by generating a shared latent content space. Hu et al. [15] propose a neural generative model that combines variational auto-encoders (VAEs) and holistic attribute discriminators, to generate sentences while controlling the attributes. Combined with the above topics, in recent years, researchers begin to focus on stylized image captioning. Gan et al. and Mathews et al. propose StyleNet [9] and SentiCap [28] to generate image captions with specific styles and sentiments, respectively. Along the same direction, we propose a novel stylized image captioning model that achieves promising performance on both tasks.

3 Method

In this section, we formally present our stylized image captioning model. Specifically, we introduce the basic encoder-decoder image captioning model in Section 3.1. In Section 3.2, we present style-factual LSTM as the core building

block of our framework. In Section 3.3, we present the overall learning strategy of the style-factual LSTM and in Section 3.4, we describe an adaptive learning approach to help the model generate stylized captions without deviating from the related image content.

3.1 Encoder-decoder Image Captioning Model

We first describe the basic encoder-decoder model [42] for image caption generation. Giving an image I and its corresponding caption $\mathbf{y} = \{y_1, ..., y_T\}$, the encoder-decoder model minimizes the following maximum likelihood estimation (MLE) loss function:

$$\theta^* = \arg\min_{\theta} \sum_{I,\mathbf{y}} \log p(\mathbf{y}|I;\theta) \tag{1}$$

where θ denotes the parameters of the model. By applying chain rule, the log likelihood of the joint probability distribution can be expressed as follows:

$$\log p(\mathbf{y}) = \sum_{t=1}^{T} \log p(y_t|y_1, ..., y_{t-1}, I) \tag{2}$$

where we drop the dependency on θ for convenience.

For the encoder-decoder image captioning model, LSTM is commonly used to model $p(y_t|y_1, ..., y_{t-1}, I)$. Specifically, it can be expressed as:

$$\begin{aligned} p(y_{t+1}|y_1, ..., y_t, I) &= f(h_t) \\ h_t &= LSTM(x_t, h_{t-1}) \end{aligned} \tag{3}$$

where h_t is the hidden state of LSTM at time t, $f(\cdot)$ is a non-linear sub-network which maps h_t into word probability distribution. For $t > 0$, x_t is the word embedding feature of word y_t; for $t = 0$, x_0 is the image feature of I.

3.2 Style-factual LSTM

To make our model capable of generating a stylized caption consistent with the image content, we devise the style-factual LSTM, which feeds two new groups of matrices $S_{x\cdot}$ and $S_{h\cdot}$ as the counterparts of $W_{x\cdot}$ and $W_{h\cdot}$, to learn to stylize the caption. In addition, at time step t, adaptive weights g_{xt} and g_{ht} are synchronously learned to adjust the relative attention weights between $W_{x\cdot}$ and $S_{x\cdot}$ as well as $W_{h\cdot}$ and $S_{h\cdot}$. The structure of style-factual LSTM is shown as Figure 2. In particular, the style-factual LSTM are defined as follows:

$$\begin{aligned} i_t &= \sigma((g_{xt}S_{xi} + (1-g_{xt})W_{xi})x_t + (g_{ht}S_{hi} + (1-g_{ht})W_{hi})h_{t-1} + b_i) \\ f_t &= \sigma((g_{xt}S_{xf} + (1-g_{xt})W_{xf})x_t + (g_{ht}S_{hf} + (1-g_{ht})W_{hf})h_{t-1} + b_f) \\ o_t &= \sigma((g_{xt}S_{xo} + (1-g_{xt})W_{xo})x_t + (g_{ht}S_{ho} + (1-g_{ht})W_{ho})h_{t-1} + b_o) \\ \widetilde{c}_t &= \phi((g_{xt}S_{xc} + (1-g_{xt})W_{xc})x_t + (g_{ht}S_{hc} + (1-g_{ht})W_{hc})h_{t-1} + b_c) \\ c_t &= f_t \odot c_{t-1} + i_t \odot \widetilde{c}_t \\ h_t &= o_t \odot \phi(c_t) \end{aligned} \tag{4}$$

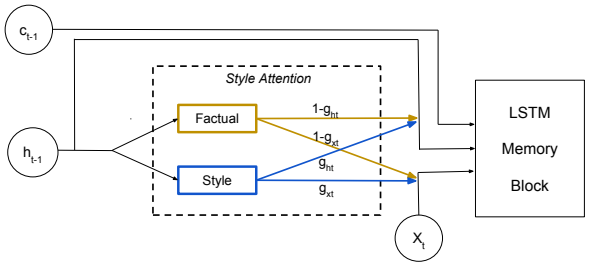

Fig. 2. An illustration of the style-factual LSTM block. Four weights, $1 - g_{ht}$, $1 - g_{xt}$, g_{ht} and g_{xt}, are designed to control the proportions of W_{hi}, W_{xi}, S_{hi} and S_{xi} matrices, respectively.

where $W_{x\cdot}$ and $W_{h\cdot}$ are responsible for generating the factual caption based on the input image, while $S_{x\cdot}$ and $S_{h\cdot}$ are responsible for adding specific style into the caption. At time step t, the style-factual LSTM feeds h_{t-1} into two independent sub-networks with one output node, which in the end figures out g_{xt} and g_{ht} after using the *sigmoid* unit to map the outputs to the range of (0, 1). Intuitively, when the model aims to predict a factual word, g_{xt} and g_{ht} should be close to 0, which encourages the model to predict the word based on $W_{x\cdot}$ and $W_{h\cdot}$. On the other hand, when the model focuses on predicting a stylized word, g_{xt} and g_{ht} should be close to 1, which encourages the model to predict the word based on $S_{x\cdot}$ and $S_{h\cdot}$.

3.3 Overall Learning Strategy

Similar to [9][25], we adopt a two-stage learning strategy to train our model. For each epoch, our model is sequentially trained by two independent stages. In the first stage, we manually fix g_{xt} and g_{ht} to 0, freezing the style-related matrices $S_{x\cdot}$ and $S_{h\cdot}$. We train the model using the paired images and ground truth factual captions. In accordance with [42], for an image-caption pair, we first extract the deep-level feature of the image using a pre-trained CNN, and then map it into an appropriate space by a linear transformation matrix. For each word, we embed its corresponding one-hot vector by a word embedding layer such that each word embedding feature has the same dimension as the transformed image feature. During training, the image feature is only fed into the LSTM as an input at the first time step. In this stage, for the style-factual LSTM, only $W_{x\cdot}$ and $W_{h\cdot}$ are updated with other layers' parameters so that they focus on generating factual captions without styles. As mentioned in Section 3.1, the MLE loss is used to train the model.

In the second stage, g_{xt} and g_{ht} are learned by the two attention sub-networks mentioned in Section 3.2, as this activates $S_{x\cdot}$ and $S_{h\cdot}$ to participate in generating the stylized caption. For this stage, we use the paired images and ground truth stylized captions to train our model. In particular, different from the first stage, we update $S_{x\cdot}$ and $S_{h\cdot}$ for style-factual LSTM, with $W_{x\cdot}$ and $W_{h\cdot}$ fixed. Also,

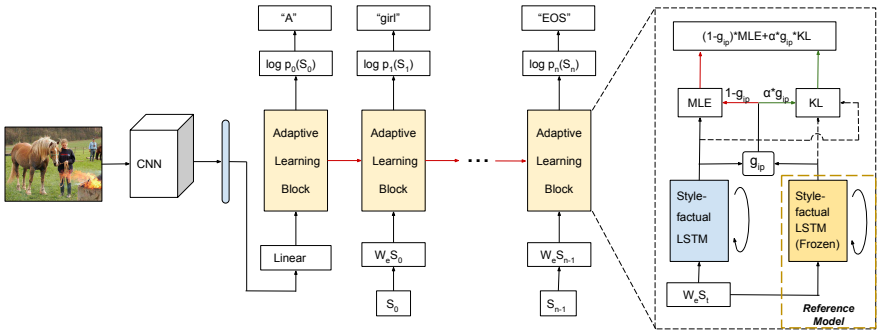

Fig. 3. The framework of our stylized image captioning model. In the adaptive learning block, the style-related matrices in the reference model (yellow) are frozen. It is designed to lead the real style-factual LSTM (blue) to learn from factual information selectively.

the parameters of the two attention sub-networks are updated concurrently with the whole network. Instead of only using the MLE loss, in Section 3.4, we will propose a novel approach to training our model in this stage.

For the test stage, to generate a stylized caption based on an image, we still compute g_{xt} and g_{ht} by the attention sub-networks, which activates $S_{x\cdot}$ and $S_{h\cdot}$. The classical beam search approach is used to predict the caption.

3.4 Adaptive Learning with Reference Factual Model

Our goal is to generate stylized captions that can accurately describe the image at the same time. Considering our style-factual LSTM, if we directly use the MLE loss to update $S_{x\cdot}$ and $S_{h\cdot}$ based on Section 3.3, it will only be updated via a few ground truth stylized captions, without learning anything from the much more massive ground truth factual captions. This may lead to the situation where the generated stylized caption cannot describe the images well. Intuitively, in a specific time step, when the generated word is unrelated to style, we encourage the model to learn more from the ground truth factual captions, instead of just a small number of the ground truth stylized captions.

Motivated by this consideration, we propose an adaptive learning approach, for which the model concurrently learns information from the ground truth stylized captions and the reference factual model, and adaptively adjusts their relative learning strengths.

In the second stage of the training process, giving an image and the corresponding ground truth stylized caption, in addition to predicting the stylized caption by the real model as Section 3.3, the framework also gives the predicted "factual version" output based on the reference model. Specifically, for reference model, we set g_{xt} and g_{ht} to 0, which freezes $S_{x\cdot}$ and $S_{h\cdot}$ as the first training stage, so that the reference model will generate its output based on $W_{x\cdot}$ and $W_{h\cdot}$. Noted that $W_{x\cdot}$ and $W_{h\cdot}$ are trained by the ground truth factual captions.

At time step t, denote the predicted word probability distribution by the real model as P_s^t, and the predicted word probability distribution by the reference model as P_r^t, we first compute their KullbackLeibler divergence (KL-divergence) as follows:

$$D(P_s^t||P_r^t) = \sum_{w \in W} P_s^t(w) \log \frac{P_s^t(w)}{P_r^t(w)} \tag{5}$$

where W is the word vocabulary. Intuitively, if the model focuses on generating a factual word, we aim to decrease $D(P_s^t||P_r^t)$, which makes P_s^t similar to P_r^t. In contrast, if the model focuses on generating a stylized word, we update the model by the MLE loss based on the corresponding ground truth stylized word.

To judge whether the current predicted word is related to style or not, we compute the inner product of P_s^t and P_r^t as the factual strength of the predicted word, we denote it as g_{ip}^t, and use it to adjust the weight between MLE and KL-divergence losses. In essence, g_{ip}^t represents the similarity between the word probability distributions P_s^t and P_r^t. When g_{ip}^t is close to 0, P_s^t has a higher possibility to correspond to a stylized word, because the reference model does not have the capacity to generate stylized words, which in the end makes g_{ip}^t small. In this situation, a higher attention weight should be given to the MLE loss. On the other hand, when g_{ip}^t is large, P_s^t has a higher possibility to correspond to a factual word, we then give KL-divergence losses higher significance.

The complete framework with the proposed adaptive learning approach is shown in Figure 3. In the end, the new loss function for the second training stage is expressed as follows:

$$Loss = \sum_{t=1}^{T} -(1 - g_{ip}^t) log P_s^t(y_t) + \alpha \cdot \sum_{t=1}^{T} g_{ip}^t D(P_s^t||P_r^t) \tag{6}$$

where α is a hyper-parameter to control the relative importance of the two loss terms. In the training process, g_{ip}^t and P_r^t do not participate in the back propagation. Still, for the style-factual LSTM, only $S_{x\cdot}$, $S_{h\cdot}$ and parameters of two attention sub-networks are updated.

4 Experiments

We perform extensive experiments to evaluate the proposed models. Experiments are evaluated by standard image captioning measurements – BLEU, Meteor, Rouge-L and CIDEr. We will first discuss the datasets and model settings used in the experiments. We then compare and analyze the results of the proposed model with the state-of-the-art stylized image captioning models.

4.1 Datasets and Model Settings

At present, there are two datasets related to stylized image captioning. First, Gan et al. [9] collect a FlickrStyle10K dataset that contains 10K Flickr images

with stylized captions. It should be noted that only the 7K training set are public. In particular, for the 7K images, each image is labeled with 5 factual captions, 1 humorous caption and 1 romantic caption. We randomly select 6000 of them as the training set, and 1000 of them as the test set. For the training set, we randomly split 10% of them as the validation set to adjust the hyper-parameters. Second, Mathews et al. [28] provide an image sentiment captioning dataset based on MSCOCO images, which contains images that are labeled by positive and negative sentiment captions. The POS subset contains 2,873 positive captions and 998 images for training, and another 2,019 captions over 673 images for testing. The NEG subset contains 2,468 negative captions and 997 images for training, and another 1,509 captions over 503 images for testing. Each of the test images has three positive and/or three negative captions. Following [28], on the training process, this sentiment dataset can be used with MSCOCO training set [3] of 413K+ factual sentences on 82K+ images, as the factual training set.

We extract image features by CNN. To make fair comparisons, for image sentiment captioning, we extract the 4096-dimension image feature by the second to last fully-connected layer of VGG-16 [36]. For stylized image captioning, we extract the 2048-dimension image feature by the last pooling layer of ResNet152 [12]. These settings are consistent with the corresponding works. Same as [28], we set the dimension of both word embedding feature and LSTM hidden state to 512 (this setting applys to all the proposed and baseline models in our experiments). For both style captioning and sentiment captioning, we use the Adam algorithm for model updating with a mini-batch size of 64 for both stages. The learning rate is set to 0.001. For style captioning, the hyper-parameter α mentioned in Section 3.4 is set to 1.1, for sentiment captioning, α is set to 0.9 and 1.5 for positive and negative captioning, which leads to the best performance in the validation set. Also, for style captioning, we directly input images into ResNet without normalization, which achieves better performance.

4.2 Performance on stylized image captioning Dataset

Experiment settings We first evaluate our proposed model on the style captioning dataset. Consistent with [9], following baselines are used for comparison:

- CaptionBot [40]: the commercial image captioning system released by Microsoft, which is trained on the large-scale factual image-caption pair data.
- Neural Image Caption (NIC) [42]: the standard encoder-decoder model for image captioning. It is trained by factual image-caption pairs of the training dataset and can generate factual captions.
- Fine-tuned: we first train an NIC, and then use the additional stylized image-caption pairs to update the parameters of the LSTM language model.
- StyleNet [9]: we train a StyleNet as [9]. To make fair comparisons, different from the original model that only uses stylized captions to update the parameter in the second stage, we train the model by the complete stylized image-caption pairs. It has two parallel model StyleNet(H) and StyleNet(R), which generate humorous and romantic captions, respectively .

Our goal is to generate captions that are both appropriately stylized and consistent with the image. There are no definite ways to separately measure these two aspects. To measure them comprehensively, for stylized captions generated by different models, we compute the BLEU-1,2,3,4, ROUGE, CIDEr, METEOR scores based on both the ground truth stylized captions and ground truth factual captions. High-performance on both situations will demonstrate the effectiveness of the stylized image captioning model for both requirements. Because we split the dataset in a different way, we re-implement all the models and compute the scores instead of directly citing them from [9].

Table 1. BLEU-1,2,3,4, ROUGE, CIDEr, METEOR scores of the proposed model and state-of-the-art methods based on ground truth stylized and factual references. "SF-LSTM" and "Adap" represents style-factual LSTM and adaptive learning approach.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	CIDEr	METEOR
Humorous/Factual Generations + Humorous References							
CaptionBot	19.7	9.5	5.1	2.8	22.8	28.1	8.9
NIC	25.4	13.3	7.4	4.2	24.3	34.1	10.4
Fine-tuned(H)	26.5	13.6	7.6	4.3	24.4	35.4	10.6
StyleNet(H)	24.1	11.7	6.5	3.9	22.3	30.7	9.4
SF-LSTM(H) (ours)	26.8	14.2	8.2	4.9	24.8	**39.8**	**11.0**
SF-LSTM + Adap(H) (ours)	**27.4**	**14.6**	**8.5**	**5.1**	**25.3**	39.5	**11.0**
Romantic/Factual Generations + Romantic References							
CaptionBot	18.4	8.7	4.5	2.4	22.3	25.0	8.7
NIC	24.3	12.8	7.4	4.4	24.1	33.7	10.2
Fine-tuned(R)	26.8	13.6	7.7	4.6	24.8	36.6	11.0
StyleNet(R)	25.4	11.7	6.1	3.5	23.2	27.9	10.0
SF-LSTM(R) (ours)	27.4	14.2	8.1	**4.9**	25.0	37.4	11.1
SF-LSTM + Adap(R) (ours)	**27.8**	**14.4**	**8.2**	4.8	**25.5**	**37.5**	**11.2**
Humorous Generations + Factual References							
Fine-tuned(H)	48.0	31.1	19.9	12.6	39.5	26.2	18.1
StyleNet(H)	45.8	28.5	17.6	11.3	36.3	22.7	16.3
SF-LSTM(H) (ours)	47.8	31.7	20.6	13.1	39.8	28.2	18.7
SF-LSTM + Adap(H) (ours)	**51.5**	**34.6**	**23.1**	**15.4**	**41.7**	**34.2**	**19.3**
Romantic Generations + Factual References							
Fine-tuned(R)	46.4	30.4	20.2	13.5	38.5	24.0	18.2
StyleNet(R)	44.2	26.8	16.3	10.4	35.4	15.8	16.3
SF-LSTM(R) (ours)	47.1	30.5	19.8	12.8	38.8	23.5	18.4
SF-LSTM + Adap(R) (ours)	**48.2**	**31.5**	**20.6**	**13.5**	**40.2**	**26.7**	**18.7**

Experiment results Table 1 shows the quantitative results of different models based on different types of ground truth captions. Considering that for each image of the test set, we only have one ground truth stylized caption instead of five, excepts CIDEr, the overall performance of other measures based on the ground truth stylized captions is reasonably lower than [9], because these measures are sensitive to the number of ground truth captions of each image. From the results, we can see that our proposed model achieves the best performance by almost all measures, regardless of testing on stylized or factual references. This demonstrates the effectiveness of our proposed model. In addition, we could see

that feeding adaptive learning approach into our model can remarkably improve the scores based on factual references, for both humorous and romantic caption generations. This indicates the improvement for generated captions' affinity toward the images. Compared with directly training the model by stylized references using MLE loss, adaptive learning can guide the model to preserve factual information in a better way, when it focuses on generating a non-stylized word.

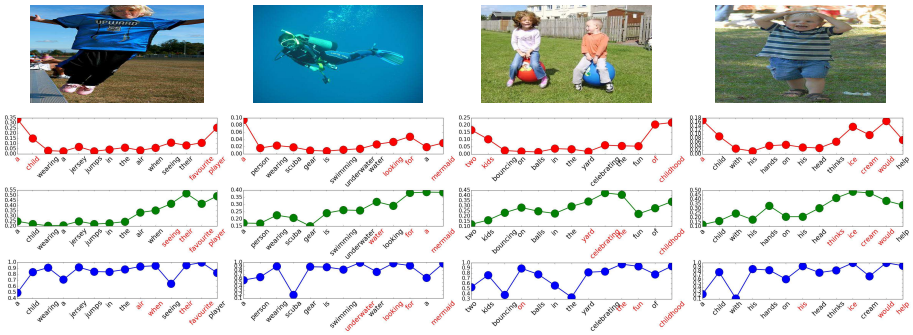

Fig. 4. Visualization of g_{xt}, g_{ht} and $1 - g_{ip}$ on several examples. The *second*, *third* and *fourth* rows correspond to g_{xt}, g_{ht}. and $1 - g_{ip}$, respectively. The *first* row is the input image. The X-axis shows the ground truth output words and the Y-axis is the weight. The top-4 words with the highest scores are in red color.

In order to prove that the proposed model is effective, we visualize the attention weights of g_{xt}, g_{ht} and $1 - g_{ip}$ mentioned in Section 3 on several examples. Specifically, we directly input the ground truth stylized caption into the trained model step by step, so that at each time step, the model will give a predicted word based on the current input word and previous hidden state. This setting simulates the training process. For each time step, Figure 4 shows the ground truth output word and the corresponding attention weights. From the first example, we could see that when the model aims to predict stylized words, "seeing", "their", "favourite", "player", g_{xt} (red line) and g_{ht} (green line) increase remarkably, indicating that when the model predicts these words, it pays more attention to the $S_{x\cdot}$ and $S_{h\cdot}$ matrices, which capture the stylized information. Otherwise, it will focus more on $W_{x\cdot}$ and $W_{h\cdot}$, which are learned to generate factual words. On the other hand, from the fourth row, when it aims to generate words "air", "when", "their", "favourite", the predicted word probability distribution similarity between the real and reference models is very low, this encourages the model to directly learn to generate these words by the MLE loss. Otherwise, it will pay considerable attention to the output of the reference model, which contains knowledge learned from ground truth factual captions. For the other three examples, still, when generating stylized phrases (i.e. "looking for a me", "celebrating the fun of childhood" and "thinks ice cream help"), overall, the style-factual LSTM can effectively give more attention to $S_{x\cdot}$ and $S_{h\cdot}$, such that

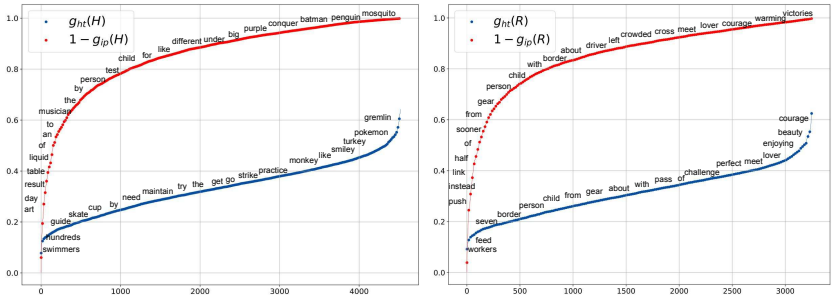

Fig. 5. The Mean value of $1 - g_{ip}$ and g_{ht} for different words. *Left*: humorous words. *Right*: romantic words.

Fig. 6. Examples of stylized captions generated by our model for different images.

it will be trained mostly by corresponding ground truth words. When generating non-stylized words, the model will focus more on the factual part in the training and predicting process. It should be noticed that the first word always gets a relative high value for g_{xt}. This is reasonable because it is usually the same word (i.e. "a") for both factual and stylized captions, the model thus cannot learn to give more attention to fact-related matrices at this very beginning. Also, some articles and prepositions, such as "a", "of", has low $1 - g_{ip}$ even if they belong to a stylized phrase. This is also reasonable and acceptable, because both the real model and reference model can predict it, there is no need to pay all the attention to the corresponding ground truth stylized word.

To further substantiate that our model successfully differentiates between stylized words and factual words, following the visualization process, we compute the mean value of $1 - g_{ip}$ and g_{ht} for each word in stylized dataset. As Figure 5 shows, words that appear frequently in the stylized parts but rarely in the factual parts tend to get higher g_{ht}. Such as "gremlin", "pokeman", "smiley" in humorous sentences and "courage", "beauty", "lover" in romantic sentences. Words that appear in the stylized and factual parts with similar frequencies are likely to hold neutral value, such as "with", "go", "of", "about". Words such as "swimmer", "person", "skate", "cup", which appear mostly in the factual parts rather than the stylized parts, tend to have lower g_{ht} scores. Since g_{ht} represents the stylized weights in the style-factual LSTM, the result of g_{ht} substantiates that the style-factual LSTM is able to differentiate between stylized and factual words. When it comes to $1 - g_{ip}$, the first kind of words we mentioned above

still receive high scores. However, we do not observe any clear border between the second and third kinds of words as g_{ht} shows. Still, we attribute it to the fact that predicting a factual noun is overall more difficult than predicting an article or preposition, which makes its corresponding inner product lower, and thus makes $1 - g_{ip}$ higher.

To make our discussion more intuitive, we show several stylized captions generated by our model in Figure 6. As Figure 6 shows, our model can generate stylized captions that accurately describe the corresponding images. For different images, the generated captions contain appropriate humorous phrases like "reach outer space","catch bones","like a lizard" and appropriate romantic phrases like "to meet his lover","speed to finish the line","conquer the high".

4.3 Performance on Image Sentiment Captioning Dataset

Table 2. BLEU-1,2,3,4, ROUGE, CIDEr, METEOR scores of the proposed model and the state-of-the-art methods for sentiment captioning.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	CIDEr	METEOR
POS Test Set							
NIC	48.7	28.1	17.0	10.7	36.6	55.6	15.3
ANP-Replace	48.2	27.8	16.4	10.1	36.6	55.2	16.5
ANP-Scoring	48.3	27.9	16.6	10.1	36.5	55.4	16.6
LSTM-Transfer	49.3	29.5	17.9	10.9	37.2	54.1	**17.0**
SentiCap	49.1	29.1	17.5	10.8	36.5	54.4	16.8
SF-LSTM + Adap (ours)	**50.5**	**30.8**	**19.1**	**12.1**	**38.0**	**60.0**	16.6
NEG Test Set							
NIC	47.6	27.5	16.3	9.8	36.1	54.6	15.0
ANP-Replace	48.1	28.8	17.7	10.9	36.3	56.5	16.0
ANP-Scoring	47.9	28.7	17.7	11.1	36.2	57.1	16.0
LSTM-Transfer	47.8	29.0	18.7	12.1	36.7	55.9	16.2
SentiCap	50.0	**31.2**	**20.3**	13.1	37.9	**61.8**	**16.8**
SF-LSTM + Adap (ours)	**50.3**	31.0	20.1	**13.3**	**38.0**	59.7	16.2

We also evaluate our model on the image sentiment caption dataset which is collected by [28]. Following [28], we compare the proposed model with several baselines. Besides NIC, ANP-Replace is based on NIC. For each caption generated by NIC, it randomly chooses a noun and adds the most common adjective of the corresponding sentiment for the chosen noun. In a similar way, ANP-Scoring uses multi-class logistic regression to select the most likely adjective for the chosen noun. LSTM-Transfer earns a fine-tuned LSTM from the sentiment dataset with additional regularization as [34]. Senticap implements a switching LSTM with word-level regularization to generate stylized captions. It should be mentioned that Senticap utilizes ground truth word sentiment strength in their regularization, which are labeled by humans. In contrast, our model only needs ground truth image-caption pairs without extra information.

Table 2 shows the performance of different models on the sentiment captioning dataset. The performance score of all baselines are directly cited from [28].

Fig. 7. Examples of sentiment caption generation based on our model. Positive and negative words are highlighted in red and blue colors.

We can see that for positive caption generation, the performance of our proposed model remarkably outperforms other baselines, with the highest scores by almost all measures. For negative caption generation, the performance of our model is competitive with Senticap while outperforming all others. Overall, without using extra ground truth information, our model achieves the best performance for generating image captions with sentiment. Figure. 7 illustrates several sentiment captions generated by our model, as it can effectively generate captions with the sentiment elements being specified.

5 Conclusions

In this paper, we present a new stylized image captioning model. We design a style-factual LSTM as the core building block of the model, which feeds two groups of matrices into the LSTM to capture both factual and stylized information. To allow the model to preserve factual information in a better way, we leverage the reference model and develop an adaptive learning approach to adaptively adding factual information into the model, based on the prediction similarity between the real and reference models. Experiments on two stylized image captioning datasets demonstrate the effectiveness of our proposed approach. It outperforms the state-of-the-art models for stylized image captioning without using extra ground truth information. Furthermore, visualization of different attention weights demonstrates that our model can indeed differentiate the factual part and stylized part of a caption automatically, and adjust the attention weights adaptively for better learning and prediction.

6 Acknowledgment

We would like to thank the support of New York State through the Goergen Institute for Data Science, our corporate sponsor Adobe and NSF Award #1704309.

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and vqa. arXiv preprint arXiv:1707.07998 (2017)
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
3. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)
4. Chen, X., Lawrence Zitnick, C.: Mind's eye: A recurrent visual representation for image caption generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2422–2431 (2015)
5. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2625–2634 (2015)
6. Elliott, D., Keller, F.: Image description using visual dependency representations. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 1292–1302 (2013)
7. Fang, H., Gupta, S., Iandola, F., Srivastava, R., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J., et al.: From captions to visual concepts and back (2015)
8. Farhadi, A., Hejrati, M., Sadeghi, M.A., Young, P., Rashtchian, C., Hockenmaier, J., Forsyth, D.: Every picture tells a story: Generating sentences from images. In: European conference on computer vision. pp. 15–29. Springer (2010)
9. Gan, C., Gan, Z., He, X., Gao, J., Deng, L.: Stylenet: Generating attractive visual captions with styles. In: CVPR (2017)
10. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on. pp. 2414–2423. IEEE (2016)
11. Gurari, D., Li, Q., Stangl, A.J., Guo, A., Lin, C., Grauman, K., Luo, J., Bigham, J.P.: Vizwiz grand challenge: Answering visual questions from blind people. arXiv preprint arXiv:1802.08218 (2018)
12. 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. pp. 770–778 (2016)
13. Hermann, K.M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., Blunsom, P.: Teaching machines to read and comprehend. In: Advances in Neural Information Processing Systems. pp. 1693–1701 (2015)
14. Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research **47**, 853–899 (2013)
15. Hu, Z., Yang, Z., Liang, X., Salakhutdinov, R., Xing, E.P.: Toward controlled generation of text. In: International Conference on Machine Learning. pp. 1587–1596 (2017)
16. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision. pp. 694–711. Springer (2016)
17. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3128–3137 (2015)

18. Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A.C., Berg, T.L.: Baby talk: Understanding and generating image descriptions. In: Proceedings of the 24th CVPR. Citeseer (2011)
19. Kuznetsova, P., Ordonez, V., Berg, A.C., Berg, T.L., Choi, Y.: Collective generation of natural image descriptions. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. pp. 359–368. Association for Computational Linguistics (2012)
20. Kuznetsova, P., Ordonez, V., Berg, T., Choi, Y.: Treetalk: Composition and compression of trees for image descriptions. Transactions of the Association of Computational Linguistics **2**(1), 351–362 (2014)
21. Lebret, R., Pinheiro, P.O., Collobert, R.: Simple image description generator via a linear phrase-based approach. arXiv preprint arXiv:1412.8419 (2014)
22. Li, S., Kulkarni, G., Berg, T.L., Berg, A.C., Choi, Y.: Composing simple image descriptions using web-scale n-grams. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning. pp. 220–228. Association for Computational Linguistics (2011)
23. Li, Y., Yao, T., Mei, T., Chao, H., Rui, Y.: Share-and-chat: Achieving human-level video commenting by search and multi-view embedding. In: Proceedings of the 2016 ACM on Multimedia Conference. pp. 928–937. ACM (2016)
24. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). vol. 6 (2017)
25. Luong, M.T., Le, Q.V., Sutskever, I., Vinyals, O., Kaiser, L.: Multi-task sequence to sequence learning. arXiv preprint arXiv:1511.06114 (2015)
26. Mao, J., Wei, X., Yang, Y., Wang, J., Huang, Z., Yuille, A.L.: Learning like a child: Fast novel visual concept learning from sentence descriptions of images. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2533–2541 (2015)
27. Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., Yuille, A.: Deep captioning with multimodal recurrent neural networks (m-rnn). arXiv preprint arXiv:1412.6632 (2014)
28. Mathews, A.P., Xie, L., He, X.: Senticap: Generating image descriptions with sentiments. In: AAAI. pp. 3574–3580 (2016)
29. Mnih, V., Heess, N., Graves, A., et al.: Recurrent models of visual attention. In: Advances in neural information processing systems. pp. 2204–2212 (2014)
30. Neumann, L., Neumann, A.: Color style transfer techniques using hue, lightness and saturation histogram matching. In: Computational Aesthetics. pp. 111–122. Citeseer (2005)
31. Ordonez, V., Kulkarni, G., Berg, T.L.: Im2text: Describing images using 1 million captioned photographs. In: Advances in neural information processing systems. pp. 1143–1151 (2011)
32. Rocktäschel, T., Grefenstette, E., Hermann, K.M., Kočiskỳ, T., Blunsom, P.: Reasoning about entailment with neural attention. arXiv preprint arXiv:1509.06664 (2015)
33. Rush, A.M., Chopra, S., Weston, J.: A neural attention model for abstractive sentence summarization. arXiv preprint arXiv:1509.00685 (2015)
34. Schweikert, G., Rätsch, G., Widmer, C., Schölkopf, B.: An empirical analysis of domain adaptation algorithms for genomic sequence analysis. In: Advances in Neural Information Processing Systems. pp. 1433–1440 (2009)

35. Shen, T., Lei, T., Barzilay, R., Jaakkola, T.: Style transfer from non-parallel text by cross-alignment. In: Advances in Neural Information Processing Systems. pp. 6833–6844 (2017)
36. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
37. Spratling, M.W., Johnson, M.H.: A feedback model of visual attention. Journal of cognitive neuroscience **16**(2), 219–237 (2004)
38. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. pp. 3104–3112 (2014)
39. Tang, Y., Srivastava, N., Salakhutdinov, R.R.: Learning generative models with visual attention. In: Advances in Neural Information Processing Systems. pp. 1808–1816 (2014)
40. Tran, K., He, X., Zhang, L., Sun, J.: Rich image captioning in the wild. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2016 IEEE Conference on. pp. 434–441. IEEE (2016)
41. Ulyanov, D., Lebedev, V., Vedaldi, A., Lempitsky, V.S.: Texture networks: Feed-forward synthesis of textures and stylized images. In: ICML. pp. 1349–1357 (2016)
42. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on. pp. 3156–3164. IEEE (2015)
43. Wu, Z.Y.Y.Y.Y., Cohen, R.S.W.W.: Encode, review, and decode: Reviewer module for caption generation. arXiv preprint arXiv:1605.07912 (2016)
44. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning. pp. 2048–2057 (2015)
45. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4651–4659 (2016)