Towards Generating Stylized Image Captions via Adversarial Training

Omid Mohamad Nezami[1,2](✉), Mark Dras[1], Stephen Wan[2],
Cécile Paris[1,2], and Len Hamey[1]

[1] Macquarie University, Sydney, NSW, Australia
omid.mohamad-nezami@hdr.mq.edu.au
{mark.dras,len.hamey}@mq.edu.au
[2] CSIRO's Data61, Sydney, NSW, Australia
{stephen.wan,cecile.paris}@data61.csiro.au

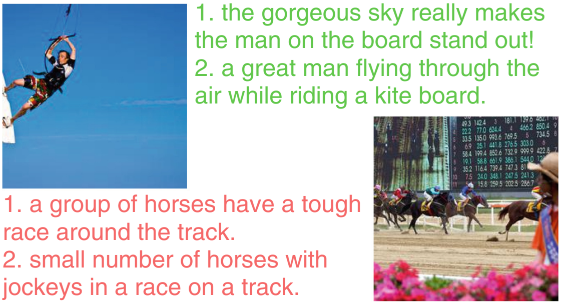
Abstract. While most image captioning aims to generate objective descriptions of images, the last few years have seen work on generating visually grounded image captions which have a specific style (e.g., incorporating positive or negative sentiment). However, because the stylistic component is typically the last part of training, current models usually pay more attention to the style at the expense of accurate content description. In addition, there is a lack of variability in terms of the stylistic aspects. To address these issues, we propose an image captioning model called ATTEND-GAN which has two core components: first, an attention-based caption generator to strongly correlate different parts of an image with different parts of a caption; and second, an adversarial training mechanism to assist the caption generator to add diverse stylistic components to the generated captions. Because of these components, ATTEND-GAN can generate correlated captions as well as more human-like variability of stylistic patterns. Our system outperforms the state-of-the-art as well as a collection of our baseline models. A linguistic analysis of the generated captions demonstrates that captions generated using ATTEND-GAN have a wider range of stylistic adjectives and adjective-noun pairs.

Keywords: Image captioning · Attention mechanism · Adversarial training

1 Introduction

Deep learning has facilitated the task of supplying images with captions. Current image captioning models [2,27,29] have gained considerable success due to powerful deep learning architectures and large image-caption datasets including the MSCOCO dataset [17]. These models mostly aim to describe an image in a factual way. Humans, however, describe an image in a way that combines subjective and stylistic properties, such as positive and negative sentiment, as in

A. C. Nayak and A. Sharma (Eds.): PRICAI 2019, LNAI 11670, pp. 270–284, 2019.
https://doi.org/10.1007/978-3-030-29908-8_22



1. the gorgeous sky really makes the man on the board stand out!
2. a great man flying through the air while riding a kite board.

1. a group of horses have a tough race around the track.
2. small number of horses with jockeys in a race on a track.

Fig. 1. Examples of positive (green) and negative (red) captions. (Color figure online)

the captions of Fig. 1. Users often find such captions more expressive and more attractive [8]; they have the practical purpose of enhancing the engagement level of users in social applications (e.g., chatbots) [14], and can assist people to make interesting image captions in social media content [8]. Moreover, Mathews *et al.* [19] found that they are more common in the descriptions of online images, and can have a role in transferring visual content clearly [18].

In stylistically enhanced descriptions, the content of images should still be reflected correctly. Moreover, the descriptions should fluently include stylistic words or phrases. To meet these criteria, previous models have used two-stage training: first, training on a large factual dataset to describe the content of an image; and then training on a small stylistic dataset to apply stylistic properties to a caption. The models have different strategies for integrating the learned information from the datasets. SentiCap has two Long Short-Term Memory (LSTM) networks: one learns from a factual dataset and the other one learns from a stylistic dataset [19]. In comparison, Gan *et al.* [8] proposed a new type of LSTM network, factored LSTM, to learn both factual and stylistic information. The factored LSTM has three matrices instead of one multiplied to the input caption: two matrices are learned to preserve the factual aspect of the input caption and one is learned to transfer the style aspect of the input caption. Chen *et al.* [5] applied an attention-based model which is similar to the factored LSTM, but it has an attention mechanism to differentiate attending to the factual and sentiment information of the input caption.

However, since the stylistic dataset is usually small, preserving the correlations between images and captions as well as generating a wide variety of stylistic patterns is very difficult. An imperfect caption from the system of Mathews *et al.* [19]—"a dead man doing a clever trick on a skateboard at a skate park"—illustrates the problem: the man is not actually dead; this is just a frequently used negative adjective.

Recently, Mathews *et al.* [18] dealt with this by applying a large stylistic dataset to separate the semantic and stylistic aspects of the generated captions. However, evaluation in this work was more difficult because the dataset includes stylistic captions which are not aligned to images. To address this challenge without any large stylistic dataset, we propose ATTEND-GAN, an image captioning model using an attention mechanism and a Generative Adversarial Network (GAN); our particular goal is to better apply stylistic information in the sort of two-stage architecture in previous work. Similar to this previous work, we first train a caption generator on a large factual dataset, although ATTEND-GAN uses an attention-based version attending to different image regions in the caption generation process [2]. Because of this, each word of a generated caption is conditioned upon a relevant fine-grained region of the corresponding image, ensuring a direct correlation between the caption and the image. Then we train a caption discriminator to distinguish between captions generated by our caption generator, and real captions, generated by humans. In the next step, on a small stylistic dataset, we implement an adversarial training mechanism to guide the generator to generate sentiment-bearing captions. To do so, the generator is trained to fool the discriminator by generating correlated and highly diversified captions similar to human-generated ones. The discriminator also periodically improves itself to further challenge the generator. Because GANs are originally designed to face continuous data distributions not discrete ones like texts [9], we use a gradient policy [31] to guide our caption generator using the rewards received from our caption discriminator for the next generated word, as in reinforcement learning [23]. The contributions of this paper are[1]:

- To generate human-like stylistic captions in a two-stage architecture, we propose ATTEND-GAN (Sect. 3) using both the designed attention-based caption generator and the adversarial training mechanism [9].
- ATTEND-GAN achieves results which are significantly better than the state-of-the-art (Sect. 4.5) and a comprehensive range of our baseline models (Sect. 4.6) for generating image captions with styles.
- On the SentiCap dataset [19], we show how ATTEND-GAN can result in stylistic captions which are strongly correlated with visual content (Sect. 4.8). ATTEND-GAN also exhibits significant variety in generating adjectives and adjective-noun pairs (Sect. 4.7).

2 Related Work

2.1 Image Captioning

The encoder-decoder framework of Vinyals *et al.* [27] where the encoder learns to encode visual content, using a Convolutional Neural Network (CNN), and the decoder learns to describe the visual content, using a long-short term memory (LSTM) network, is the basis of modern image captioning systems. Having an attention-based component has resulted in the most successful image captioning models [2,22,29,30]. These models use attention in either the image side or the caption side. For instance, Xu *et al.* [29] and Rennie *et al.* [22] attended to

[1] Our code and trained model are publicly available from https://github.com/omidmnezami/ATTEND-GAN.

the spatial visual features of an image. In comparison, You *et al.* [30] applied semantic attention attending to visual concepts detected in an image. Anderson *et al.* [2] applied an attention mechanism to attend to spatial visual features and discriminate not only the visual regions but also the detected concepts in the regions [2]. In addition to factual image captioning, the ability to generate stylistic image captions has recently become popular. The key published work [5,8,18,19] uses a two-stage architecture, although end-to-end is possible. None of the existing work uses an adversarial training mechanism; we show this, combined with attention, significantly outperforms the previous work.

2.2 Generative Adversarial Network

Goodfellow *et al.* [9] introduced Generative Adversarial Networks (GANs), whose training mechanism consists of a generator and a discriminator; they have been applied with great success in different applications [12,15,21,28,31]. The discriminator is trained to recognize real and synthesized samples generated by the generator. In contrast, the generator wants to generate realistic data to mislead the discriminator in distinguishing the source of data.

GANs were originally established for a continuous data space [9,31] rather than a discrete data distribution as in our work. To handle this, a form of reinforcement learning is usually applied, where the sentence generation process is formulated as a reinforcement learning problem [23]; the discriminator provides a reward for the next action (in our context the next generated word), and the generator uses the reward to calculate gradients and update its parameters, as proposed in Yu *et al.* [31]. Wang and Wan [28] applied this to generating sentiment-bearing text (although not conditioned on any input, such as the images in our captioning task).

3 ATTEND-GAN Model

The purpose of our image captioning model is to generate sentiment-bearing captions. Our caption generator employs an attention mechanism, described in

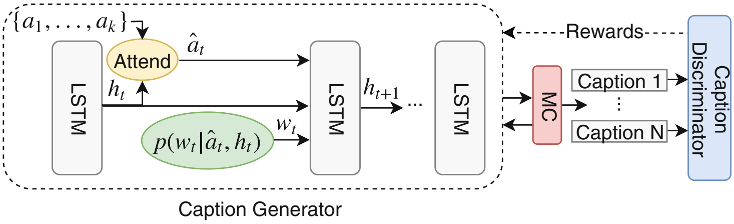


Fig. 2. The architecture of the ATTEND-GAN model. $\{a_1, ..., a_K\}$ are spatial visual features generated by ResNet-152 network. Attend and MC modules are our attention mechanism and Monte Carlo search, respectively.

Sect. 3.1, to attend to fine-grained image regions $a = \{a_1, ..., a_K\}, a_i \in \mathbb{R}^D$, where the number of regions is K with D dimensions, in different time steps so as to generate an image caption $x = \{x_1, \ldots, x_T\}, x_i \in \mathbb{R}^N$, where the size of our vocabulary is N and the length of the generated caption is T. We also propose a caption discriminator, explained in Sect. 3.2, to distinguish between the generated captions and human-produced ones. We describe our training in Sect. 3.3. Our proposed model is called ATTEND-GAN (Fig. 2).

3.1 Caption Generator

The goal of our caption generator $G_\theta(x_t|x_{1:t-1}, \hat{a}_t)$ is to generate an image caption to achieve a maximum reward value from our caption discriminator $D_\phi(x_{1:T})$, where θ and ϕ are the parameters of the generator and the discriminator, respectively. The objective function of the generator, which is dependent on the discriminator, is to minimize:

$$L_1(\theta) = \sum_{1 \le t \le T} G_\theta(x_t|x_{1:t-1}, \hat{a}_t).Z_{D_\phi}^{G_\theta}(x_{1:t}) \tag{1}$$

where $Z_{D_\phi}^{G_\theta}(x_{1:t})$ is the reward value of the partially generated sequence, $x_{1:t}$, and is estimated using the discriminator. The reward value can be interpreted as a score value that $x_{1:t}$ is real. Since the discriminator can only generate a reward value for a complete sequence, Monte Carlo (MC) search is applied, which uses the generator to roll out the remaining part of the sequence at each time step. We apply MC search N times, and calculate the average reward (to decrease the variance of the next generated words):

$$Z_{D_\phi}^{G_\theta}(x_{1:t}) = \begin{cases} \frac{1}{N} \sum\limits_{n=1}^{N} D_\phi(x_{1:T}^n), \; x_{1:T}^n \in MC_{G_\theta}(x_{1:t;N}) & \text{if } t < T \\ D_\phi(x_{1:t}) & \text{if } t = T \end{cases} \tag{2}$$

$x_{1:T}^n$ is the n-th MC-completed sequence at current time step t. In addition to Eq. (1), we calculate the maximum likelihood estimation (MLE) of the generated word with respect to the attention-based content ($\hat{a}_t$) and the hidden state (h_t) at the current time of our LSTM, which is the core of our caption generator, as the second objective function:

$$L_2(\theta) = -\sum_{1 \le t \le T} \log(p_w(x_t \,|\, \hat{a}_t, h_t)) + \lambda_1 \sum_{1 \le k \le K} (1 - \sum_{1 \le t \le T} a_{tk})^2 \tag{3}$$

p_w is calculated using a multilayer perceptron with a softmax layer on its output and indicates the probabilities of the possible generated words:

$$p_w(x_t \,|\, \hat{a}_t, h_t) = \text{softmax}(\hat{a}_t W_a + h_t W_h + b_w) \tag{4}$$

W_x and b_w are the learned weights and biases. The last term in Eq. (3) is to encourage our caption generator to equally consider diverse regions of the given

Algorithm 1. ATTEND-GAN Training Mechanism.

```
Pre-train the caption generator ($G_\theta$) using Eq. (9).
Use $G_\theta$ to generate sample captions $\mathbb{P}_G$ and select ground-truth captions $\mathbb{P}_H$.
Pre-train the caption discriminator ($D_\phi$) using Eq. (10) and the combination of $\mathbb{P}_G$ and $\mathbb{P}_H$.
repeat
    for $g$ steps do
        Apply $G_\theta$ to generate image captions.
        Calculate $Z_{D_\phi}^{G_\theta}$ using Eq. (2).
        Update $\theta$, the parameters of $G_\theta$, using Eq. (8).
    end for
    for $d$ steps do
        Generate sample captions $\mathbb{P}_G$ by $G_\theta$ and select human-generated captions $\mathbb{P}_H$.

        Update $\phi$, the parameters of $D_\phi$, using Eq. (10).
    end for
until ATTEND-GAN converges
```

image at the end of the caption generation process. λ_1 is a regularization parameter. h_t is calculated using our LSTM:

$$\begin{aligned}
i_t &= \sigma(H_i h_{t-1} + W_i w_{t-1} + A_i \hat{a}_t + b_i)\\
f_t &= \sigma(H_f h_{t-1} + W_f w_{t-1} + A_f \hat{a}_t + b_f)\\
g_t &= \tanh(H_g h_{t-1} + W_g w_{t-1} + A_g \hat{a}_t + b_g)\\
o_t &= \sigma(H_o h_{t-1} + W_o w_{t-1} + A_o \hat{a}_t + b_o)\\
c_t &= f_t c_{t-1} + i_t g_t\\
h_t &= o_t \tanh(c_t)
\end{aligned} \tag{5}$$

Here, i_t, f_t, g_t, o_t, and c_t are the LSTM's gates and represent input, forget, modulation, output, and memory gates, respectively. w_{t-1} is the embedded previous word in M dimensions, $w_x \in \mathbb{R}^M$. H_x, W_x, A_x, and b_x are learned weights and biases; and σ is the Sigmoid function. Using h_t, our soft attention module generates unnormalized weights $e_{j,t}$ for each image region a_j. Then, the weights are normalized using a softmax layer, e'_t:

$$e_{j,t} = W_e^T \tanh(W'_a a_j + W'_h h_t), e'_t = \text{softmax}(e_t) \tag{6}$$

W_e^T and W'_x are our trained weights. Finally, $\hat{a}_t$, our attention-based content, is calculated using Eq. (7):

$$\hat{a}_t = \sum_{1 \le j \le K} e'_{j,t} a_j \tag{7}$$

During the adversarial training, the objective function of the caption generator is a combination of Eqs. (1) and (3):

$$L_G(\theta) = \lambda_2 L_1(\theta) + L_2(\theta) \tag{8}$$

λ_2 is a balance parameter. The discriminator cannot be learned effectively from a random initialization of the generator; we therefore pretrain the generator with the MLE objective function:

$$L_G(\theta) = L_2(\theta) \tag{9}$$

3.2 Caption Discriminator

Our caption discriminator is inspired by the Wasserstein GAN (WGAN) [3] which is an improved version of the GAN [9]. The WGAN generates continuous values and solves the problem of the GAN generating non-continuous outputs leading to some training difficulties (e.g. vanishing gradients). The objective function of our WGAN is:

$$L_D(\phi) = \mathbb{E}_{x \sim \mathbb{P}_H}[D_\phi(x)] - \mathbb{E}_{\overline{\mathrm{x}} \sim \mathbb{P}_G}[D_\phi(\overline{\mathrm{x}})] \tag{10}$$

where ϕ are the parameters of the discriminator (D_ϕ); $\mathbb{P}_H$ is the set of the generated captions by humans; and $\mathbb{P}_G$ is the set of the generated captions by the generator. D_ϕ is implemented via a Convolutional Neural Network (CNN) that calculates the score value of the input caption. To feed a caption to our CNN model, we first embed all words in the caption into M embedding dimensions, $\{w'_1, \ldots, w'_T\}, w'_i \in \mathbb{R}^M$, and build a 2-dimensional matrix for the caption, $S \in \mathbb{R}^{T \times M}$ [31]. Our CNN model includes Convolutional (Conv.) layers with P different kernel sizes $\{k_1, \ldots, k_P\}, k_i \in \mathbb{R}^{C \times M}$, where C indicates the number of the words ($C \in [1, T]$). Applying each Conv. layer to S results a number of feature maps, $v_{ij} = k_i \otimes S_{j:j+C-1} + b_j$, where $\otimes$ is a convolution operation and b_j is a bias vector. We apply a batch normalization layer [11], and a nonlinearity, a rectified linear unit (ReLU), respectively. Then, we apply a max-pooling layer, $v^*_i = \max v_{ij}$. Finally, a fully connected layer is applied to output the score value of the caption. The weights of our CNN model are clipped to be in a compact space.

3.3 ATTEND-GAN Training

As shown in Algorithm 1, we first pre-train our caption generator for a specific number of epochs. Then, we apply the best generator model to generate sample captions. The real captions are selected from the ground truth. In Step 3, our caption discriminator is pre-trained using a combination of the generated and real captions for a specific number of epochs. Here, both the caption generator and discriminator are pre-trained on a factual dataset. In Step 4, we start our adversarial training on a sentiment-bearing dataset with positive or negative sentiment. We continue the training of the caption generator and discriminator for g-steps and d-steps, respectively. Using this mechanism, we improve both the caption generator and discriminator. Here, the caption generator applies the received rewards from the caption discriminator to update its parameters using Eq. (8).

4 Experiments

4.1 Datasets

Microsoft COCO Dataset. We use the MSCOCO image-caption dataset [17] to train our models. Specifically, we use the training set of the dataset including 82K+ images and 413K+ captions.

SentiCap Dataset. To add sentiment to the generated captions, our models are trained on the SentiCap dataset [19] including sentiment-bearing image captions. The dataset has two separate sections of sentiments: *positive* and *negative*. 2,873 captions paired with 998 images (409 captions with 174 images are for validation) are for training and 2019 captions paired with 673 images are for testing in the positive section. 2,468 captions paired with 997 images (429 captions with 174 images are for validation) are for training and 1,509 captions paired with 503 images are for testing in the negative section. We use the same training/test folds as in the previous work [5,19].

4.2 Evaluation Metrics

ATTEND-GAN is evaluated using standard image captioning metrics: METEOR [7], BLEU [20], CIDEr [26] and ROUGE-L [16]. SPICE has not previously been used in the literature; however, it is reported for future comparisons because it has shown a close correlation with human-based evaluations [1]. Larger values of these metrics indicated better results.

4.3 Models for Comparison

We first trained our models on the MSCOCO dataset to generate factual captions. Then, we trained our models on the SentiCap dataset to add sentiment properties to the generated captions. This two-stage training mechanism is similar to the training methods of [19] and [8]. The work of [5], the newest one in this domain, was also implemented in a similar way. Following this training approach makes our results directly comparable to the previous ones. Our models are compared with a range of baseline models from Mathews *et al.* [19]: CNN+RNN, which is only trained using the MSCOCO dataset; ANP-Replace, which adds the most common adjectives to a randomly chosen noun; ANP-Scoring, which applies multi-class logistic regression to select an adjective for the chosen noun; RNN-Transfer, which is CNN+RNN fine-tuned on the SentiCap dataset; and their key system SentiCap, which uses two LSTM modules to learn from factual and sentiment-bearing caption. We also compare with SF-LSTM+Adap, which applies an attention mechanism to weight factual and sentiment-based information [5]. The results of all these models in Table 1 are obtained from the corresponding references. Moreover, we first train our attention-based model only on the factual dataset MSCOCO (we name this model ATTEND-GAN$_{-SA}$). Second, we train our model additionally on the SentiCap dataset but without our caption discriminator (ATTEND-GAN$_{-A}$). Finally, we train our full model using the caption discriminator (ATTEND-GAN).

4.4 Implementation Details

Encoder. In this work, we apply ResNet-152 [10] as our visual encoder model pretrained using the ImageNet dataset [6]. In comparison with other CNN models, ResNet-152 has shown more effective results on different image-caption datasets [4]. We specifically use its Res5c layer to extract the spatial features of an image. The layer gives us $7 \times 7 \times 2048$ feature map converted to 49×2048 representing 49 semantic-based regions with 2048 dimensions.

Vocabulary. Our vocabulary has 9703 words, coming form both the MSCOCO and SentiCap datasets, for all our models. Each word is embedded into a 300 dimensional vector.

Generator and Discriminator. The size of the hidden state and the memory cell of our LSTM is set to 512. For the caption generator, we use the Adam function [13] for optimization and set the learning rate to 0.0001. We set the size of our mini-batches to 64. To optimize the caption discriminator, we use the RMSprop solver [24] and clip the weights to $[-0.01, 0.01]$. The mini-batches are fixed to 80 for the discriminator. We apply Monte Carlo search 5 times (Eq. (2)). We set λ_1 and λ_2 to 1.0 and 0.1 in Eqs. (3) and (8), respectively. During the adversarial training, we alternate between Eqs. (8) and (10) to optimize the generator and the discriminator, respectively. We particularly operate a single gradient decent phase on the generator (g steps) and 3 gradient phases (d steps) on the discriminator every time. The models are trained for 20 epochs to converge. The METEOR metric is used to select the model with the best performance on the validation sets of positive and negative datasets of SentiCap because it has a close correlation with human judgments and is less computationally expensive than SPICE which requires dependency parsing [1].

4.5 Results: Comparison with the State-of-the-Art

All models in Table 1 used the same training/test folds of the SentiCap dataset to make them comparable. In comparison with the state-of-the-art, our full model (ATTEND-GAN) achieves the best results for all image captioning metrics in both positive and negative parts of the SentiCap dataset. We report the average results to show the average improvements of our models over the state-of-the-art model. ATTEND-GAN achieved large gains of 6.15, 6.45, 3.00, and 2.95 points with respect to the best previous model using BLEU-1, ROUGE-L, CIDEr and BLEU-2 metrics, respectively. Other metrics show smaller but still positive improvements.

4.6 Results: Comparison with Our Baseline Models

Our models are compared in Table 1 in terms of image captioning metrics. ATTEND-GAN outperforms ATTEND-GAN$_{-A}$ over all metrics across both positive and negative parts of the SentiCap dataset; the discriminator is thus

Table 1. The compared performances on different sections of SentiCap and their average. BLEU-N metric is shown by B-N. (The best results are bold.)

Senti	Model	B-1	B-2	B-3	B-4	ROUGE-L	METEOR	CIDEr	SPICE
Pos	CNN+RNN	48.7	28.1	17.0	10.7	36.6	15.3	55.6	–
	ANP-Replace	48.2	27.8	16.4	10.1	36.6	16.5	55.2	–
	ANP-Scoring	48.3	27.9	16.6	10.1	36.5	16.6	55.4	–
	RNN-Transfer	49.3	29.5	17.9	10.9	37.2	17.0	54.1	–
	SentiCap	49.1	29.1	17.5	10.8	36.5	16.8	54.4	–
	SF-LSTM + Adap	50.5	30.8	19.1	12.1	38.0	16.6	60.0	–
	Ours: ATTEND-GAN$_{-SA}$	56.1	32.5	19.4	11.8	44.8	17.1	63.0	15.9
	Ours: ATTEND-GAN$_{-A}$	55.8	33.4	20.1	12.4	44.2	18.6	61.1	15.7
	Ours: ATTEND-GAN	56.9	33.6	20.3	12.5	44.3	18.8	61.6	15.9
Neg	CNN+RNN	47.6	27.5	16.3	9.8	36.1	15.0	54.6	–
	ANP-Replace	48.1	28.8	17.7	10.9	36.3	16.0	56.5	–
	ANP-Scoring	47.9	28.7	17.7	11.1	36.2	16.0	57.1	–
	RNN-Transfer	47.8	29.0	18.7	12.1	36.7	16.2	55.9	–
	SentiCap	50.0	31.2	20.3	13.1	37.9	16.8	61.8	–
	SF-LSTM + Adap	50.3	31.0	20.1	13.3	38.0	16.2	59.7	–
	Ours: ATTEND-GAN$_{-SA}$	55.4	32.4	19.4	11.9	44.4	17.0	63.4	15.6
	Ours: ATTEND-GAN$_{-A}$	54.7	32.6	20.4	12.9	43.2	17.7	60.4	16.1
	Ours: ATTEND-GAN	56.2	34.1	21.3	13.6	44.6	17.9	64.1	16.2
Avg	CNN+RNN	48.15	27.80	16.65	10.25	36.35	15.15	55.10	–
	ANP-Replace	48.15	28.30	17.05	10.50	36.45	16.25	55.85	–
	ANP-Scoring	48.10	28.30	17.15	10.60	36.35	16.30	56.25	–
	RNN-Transfer	48.55	29.25	18.30	11.50	36.95	16.60	55.00	–
	SentiCap	49.55	30.15	18.90	11.95	37.20	16.80	58.10	–
	SF-LSTM + Adap	50.40	30.90	19.60	12.70	38.00	16.40	59.85	–
	Ours: ATTEND-GAN$_{-SA}$	55.75	32.45	19.40	11.85	**44.60**	17.05	**63.20**	15.75
	Ours: ATTEND-GAN$_{-A}$	55.25	33.00	20.25	12.65	43.70	18.15	60.75	15.90
	Ours: ATTEND-GAN	**56.55**	**33.85**	**20.80**	**13.05**	44.45	**18.35**	62.85	**16.05**

an important part of the architecture. ATTEND-GAN outperforms ATTEND-GAN$_{-SA}$ for all metrics except, by a small margin, CIDEr and ROUGE-L. Recall that ATTEND-GAN$_{-SA}$ is trained only on the large MSCOCO (with many captions), and so is in a sense encouraged to have diverse captions; second-stage training for ATTEND-GAN$_{-A}$ and ATTEND-GAN leads to more focussed captions relevant to SentiCap. As CIDEr and ROUGE-L are the two recall-oriented metrics, they suffer in this two-stage process, illustrating the issue we noted in Sect. 1. The discriminator, however, removes almost all of this penalty, as well as boosting the other metrics beyond ATTEND-GAN$_{-SA}$. Furthermore, Sect. 4.7 illustrates how ATTEND-GAN$_{-SA}$ produces unsatisfactory captions in terms of sentiment.

Table 2. Entropy and Top_4 of the generated adjectives using different models.

Senti	Model	Entropy	Top$_4$
Pos	ATTEND-GAN$_{-SA}$	2.2457	93.33%
	ATTEND-GAN$_{-A}$	3.0324	72.11%
	ATTEND-GAN	3.5671	62.33%
Neg	ATTEND-GAN$_{-SA}$	2.2448	91.67%
	ATTEND-GAN$_{-A}$	4.1040	48.44%
	ATTEND-GAN	3.9562	50.51%
Avg	ATTEND-GAN$_{-SA}$	2.2453	92.50%
	ATTEND-GAN$_{-A}$	3.5682	60.28%
	ATTEND-GAN	**3.7617**	**56.42%**

Table 3. The top-10 adjectives that are generated by our models and are in the adjective-noun pairs of the SentiCap dataset.

Senti	Model	Top 10 adjectives
Pos	ATTEND-GAN$_{-SA}$	white, black, small, blue, different, little, busy, _, _, _
	ATTEND-GAN$_{-A}$	nice, beautiful, happy, busy, great, sunny, good, cute, pretty, white
	ATTEND-GAN	nice, beautiful, happy, great, good, sunny, busy, white, pretty, delicious
Neg	ATTEND-GAN$_{-SA}$	black, white, small, blue, different, tall, little, _, _, _
	ATTEND-GAN$_{-A}$	lonely, dead, broken, stupid, dirty, bad, cold, little, crazy, lazy
	ATTEND-GAN	lonely, stupid, broken, dirty, dead, cold, bad, white, crazy, little

4.7 Qualitative Results

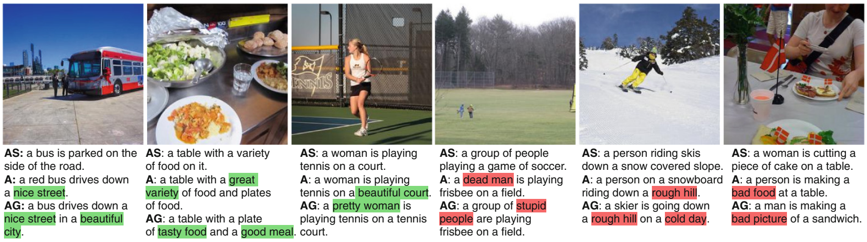
To analyze the quality of language generated by our models, we extract all generated adjectives using the Stanford part-of-speech tagger software [25], and select the adjectives found in the adjective-noun pairs (ANPs) of the SentiCap dataset. Then, we calculate Entropy of the distribution of these adjectives as a measure of variety in lexical selection (higher scores mean more variety) using Eq. (11).

$$\text{Entropy} = -\sum_{1 \leq j \leq U} \log_2[p(A_j)] \times p(A_j) \tag{11}$$

where $p(A_j)$ is the probability of the adjective (A_j) and U indicates the number of all unique adjectives. Moreover, we calculate the total probability mass of the four most frequent adjectives (Top$_4$) generated by our models. Here, lower values mean that the model allocates more probability to other generated adjectives, also indicating greater variety.

Table 2 shows that ATTEND-GAN achieves the best results on average for Entropy (highest score) and Top$_4$ (lowest) compared to other models, by a large margin with respect to ATTEND-GAN$_{-SA}$. It is not surprising that ATTEND-GAN$_{-SA}$ has the lowest variability of use of sentiment-bearing adjectives because it does not use the stylistic dataset. As demonstrated by the improvement of ATTEND-GAN over ATTEND-GAN$_{-A}$, the discriminator helps in generating a greater diversity of adjectives.

The top-10 adjectives generated by our models are shown in Table 3. "white" is generated for both negative and positive sections because they are common in both sections. ATTEND-GAN and ATTEND-GAN$_{-A}$ produce a natural ranking of sentiment-bearing adjectives for both sections. For example, these models rank "nice" as the most positive adjective, and "lonely" as the most negative. As ATTEND-GAN$_{-SA}$ does not use the stylistic dataset, it generates a similar and limited (<10) range of adjectives for both.



AS: a bus is parked on the side of the road.
A: a red bus drives down a nice street.
AG: a bus drives down a nice street in a beautiful city.

AS: a table with a variety of food on it.
A: a table with a great variety of food and plates of food.
AG: a table with a plate of tasty food and a good meal.

AS: a woman is playing tennis on a court.
A: a woman is playing tennis on a beautiful court.
AG: a pretty woman is playing tennis on a tennis court.

AS: a group of people playing a game of soccer.
A: a dead man is playing frisbee on a field.
AG: a group of stupid people are playing frisbee on a field.

AS: a person riding skis down a snow covered slope.
A: a person on a snowboard riding down a rough hill.
AG: a skier is going down a rough hill on a cold day.

AS: a woman is cutting a piece of cake on a table.
A: a person is making a bad food at a table.
AG: a man is making a bad picture of a sandwich.

Fig. 3. Examples on the positive (first 3) and negative (last 3) datasets (AS for ATTEND-GAN$_{-SA}$, A for ATTEND-GAN$_{-A}$ and AG for ATTEND-GAN). Green and red colors indicate the generated positive and negative adjective-noun pairs in SentiCap, respectively. (Color figure online)

4.8 Generated Captions

Figure 3 shows sample sentiment-bearing captions generated by our models for the positive and negative sections of the SentiCap dataset.[2] For instance, for the first two images, ATTEND-GAN correctly applies positive sentiments to describe the corresponding images (e.g., "nice street", "tasty food"). Here, ATTEND-GAN$_{-A}$ also succeeds in generating captions with positive sentiments, but less well. In the third image, ATTEND-GAN uses "pretty woman" to describe the image which is better than the "beautiful court" of ATTEND-GAN$_{-A}$: for this image, all ground-truth captions have positive sentiment for the noun "girl" (e.g. "a beautiful girl is running and swinging a tennis racket"); none of them describes the noun "court" with a sentiment-bearing adjective as

[2] See a link to supplementary materials for additional samples: https://github.com/omidmnezami/ATTEND-GAN/blob/master/st.pdf.

ATTEND-GAN$_{-A}$ does. For all images, since ATTEND-GAN$_{-SA}$ is not trained using the SentiCap dataset, it does not generate any caption with sentiment. For the fourth image, ATTEND-GAN generates "a group of stupid people are playing frisbee on a field", applying "stupid people" to describe the image negatively. Here, one of the ground-truth captions exactly includes "stupid people" ("two stupid people in open field watching yellow tent blown away"). ATTEND-GAN$_{-A}$, like our flawed example from Sect. 1, refers instead inaccurately to a dead man. For the fifth image (as for the first image), ATTEND-GAN has incorporates more (appropriate) sentiment in comparison to ATTEND-GAN$_{-A}$. It generates "rough hill" and "cold day", while ATTEND-GAN$_{-A}$ only generates the former. It also uses "skier" which is more appropriate than "person". In the last image, ATTEND-GAN adds "bad picture" and ATTEND-GAN$_{-A}$ generates "bad food". One of the ground-truth captions exactly includes "bad picture".

5 Conclusion

In this paper, we proposed ATTEND-GAN, an attention-based image captioning model using an adversarial training mechanism. Our model is capable of generating stylistic captions which are strongly correlated with images and contain diverse stylistic components. ATTEND-GAN achieves the state-of-the-art performance on the SentiCap dataset. It also outperforms our baseline models and generates stylistic captions with a high level of variety. Future work includes developing ATTEND-GAN to generate a wider range of captions and developing further mechanisms to ensure compatibility with the visual content.

References

1. Anderson, P., Fernando, B., Johnson, M., Gould, S.: SPICE: semantic propositional image caption evaluation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9909, pp. 382–398. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_24
2. Anderson, P., et al.: Bottom-up and top-down attention for image captioning and visual question answering. In: CVPR, vol. 3, p. 6 (2018)
3. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein GAN (2017). arXiv preprint arXiv:1701.07875
4. Chen, L., et al.: SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6298–6306. IEEE (2017)
5. Chen, T., et al.: "Factual" or "emotional": stylized image captioning with adaptive learning and attention (2018). arXiv preprint arXiv:1807.03871
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database (2009)
7. Denkowski, M., Lavie, A.: Meteor universal: language specific translation evaluation for any target language. In: WMT, pp. 376–380 (2014)
8. Gan, C., Gan, Z., He, X., Gao, J., Deng, L.: StyleNet: generating attractive visual captions with styles. In: CVPR. IEEE (2017)

9. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
11. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift (2015). arXiv preprint arXiv:1502.03167
12. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)
13. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization (2014). arXiv preprint arXiv:1412.6980
14. 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 24th ACM International Conference on Multimedia, pp. 928–937. ACM (2016)
15. Liang, X., Hu, Z., Zhang, H., Gan, C., Xing, E.P.: Recurrent topic-transition GAN for visual paragraph generation (2017). arXiv preprint arXiv:1703.07022
16. Lin, C.Y.: Rouge: a package for automatic evaluation of summaries. In: Text Summarization Branches Out (2004)
17. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
18. 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, pp. 8591–8600 (2018)
19. Mathews, A.P., Xie, L., He, X.: SentiCap: generating image descriptions with sentiments. In: AAAI, pp. 3574–3580 (2016)
20. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: ACL, pp. 311–318. Association for Computational Linguistics (2002)
21. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks (2015). arXiv preprint arXiv:1511.06434
22. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: CVPR, vol. 1, p. 3 (2017)
23. Silver, D., et al.: Mastering the game of go with deep neural networks and tree search. Nature **529**(7587), 484 (2016)
24. Tieleman, T., Hinton, G.: Lecture 6.5-rmsprop: divide the gradient by a running average of its recent magnitude. COURSERA: Neural Netw. Mach. Learn. **4**(2), 26–31 (2012)
25. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: NAACL HLT, pp. 173–180. Association for Computational Linguistics (2003)
26. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: consensus-based image description evaluation. In: CVPR, pp. 4566–4575. IEEE (2015)
27. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. In: CVPR, pp. 3156–3164. IEEE (2015)
28. Wang, K., Wan, X.: SentiGAN: generating sentimental texts via mixture adversarial networks. In: IJCAI, pp. 4446–4452 (2018)
29. Xu, K., et al.: Show, attend and tell: neural image caption generation with visual attention. In: ICML, pp. 2048–2057 (2015)

30. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: CVPR, pp. 4651–4659. IEEE (2016)
31. Yu, L., Zhang, W., Wang, J., Yu, Y.: SeqGAN: sequence generative adversarial nets with policy gradient. In: AAAI, pp. 2852–2858 (2017)