Cross-Domain Style Synthesis for Image Captions

Zhen Wang,[1] **Yiqun Duan,** [1] **Jingya Wang** [1]

[1] UBTECH Sydney AI Centre, School of Computer Science, Faculty of Engineering,
The University of Sydney, Darlington, NSW 2008, Australia
zwan4121@uni.sydney.edu.au; duanyiquncc@gmail.com

Abstract

In this paper, we propose a transformer-based image caption network to solve cross-domain style synthesis for image caption (CDS-Cap). To present multiple styles from different domains in a single caption, we propose a multi-stage training framework to jointly use discriminators from different domains to force generator to synchronize styles in one generated caption. However, due to the discrete property of caption generation, purely adversarial training may push generated caption deviated from real data. Based on adversarial training, we further propose to utilize real training corpus with assigned pseudo labels by domain specified discriminators to perform self-training then alleviate deviations. The proposed pseudo label self-training strategy could boost the evaluation performance of synchronized multi-style caption generation. We conduct comprehensive experiments and ablation study to demonstrate the efficiency of our proposed method.

Introduction

Semantic understanding of vision scenery has been an essential and important ability for human perceiving behaviours (Lake et al. 2017). Under scenery understanding area, image caption task, which extracts semantic features from a given picture and generates text descriptions accordingly, has attracted attentions. Current image caption systems (Vinyals et al. 2015; Zhou et al. 2016; Chen et al. 2019; Yang et al. 2017; Donahue et al. 2017) mostly focus on a factual description merging essential objects presenting in the original image. However, besides understanding the essential context in certain images, generating human-like captions also requires the control of multiple attributes (eg. sentiment, style). As mentioned in (Guo et al. 2019), incorporating attributes such as personality, emotion and sentiment could make the linguistic styles of generated captions fit the real-life scenarios better.

The controllable image caption is not a brand new idea. SentiCap (Mathews, Xie, and He 2016) firstly annotated positive and negative labels on COCO image caption, and realize simple sentiment caption generation by using paralleled word-level supervision. By then, StyleNet (Gan et al. 2017), SF-LSTM (Wen et al. 2015), and Personality Caption (Shuster et al. 2019) prove that annotate parallel style label and train the parallel style captions with word-level supervision is feasible. It is noted that, the collection and annotation of stylized image caption is expensive and not transferable between each annotated datasets. MSCAP (Guo et al. 2019) proposes a semi-supervised approach to train unpaired stylized captions from different datasets by combining generative adversarial network (GAN) and traditional image caption approaches. MSCAP treat style labels from different domains as the same level and mutually exclusive, which means MSCAP normally only generate one certain style at a time (Guo et al. 2019). However, we believe that control multiple labels from different domains at the same time (eg. both positive and humorous) is a realistic but important problem for real-life scenarios.

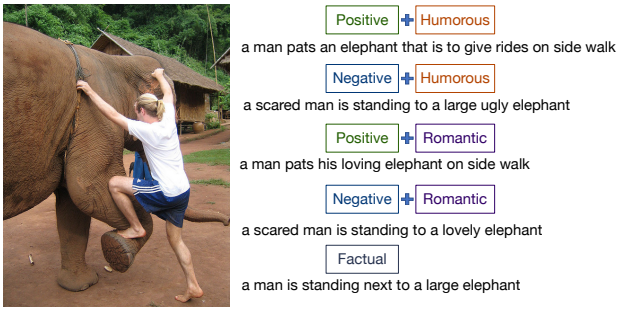

Figure 1: Example of multiple styles synthesis on a single caption

In this paper, we address the problem that given a certain image with **multiple controllable labels**, the proposed model should generate captions which present multiple styles accordingly at the same time as it shown in Fig. 1. Naturally, the best way is to manually assign multiple labels from different domains to a same caption to realize supervised parallel training. However, due to the expensiveness of collecting paired data, we propose to learn different styles from different public datasets and merge these learned styles by cross-domain alignment. For example, the model is expected to learn style A from dataset A, and synthesizes style B from dataset B. In that case, the model is expected to coherently present style A and B simultaneously. There are two crucial challenges for this compared-hard problem. 1) The image-caption-style pairs do not have intersections,

so how to solve the cross-domain style alignment is a crucial task. 2) How to merge different styles and keeps each of them effective is a crucial task.

To solve the proposed challenges mentioned above, we propose CDS Caption, an adversarial training framework for multiple-attributes controlled image caption task. Different from most previous works, we propose a transformer-based model for conditional image caption it in section . Also, instead of using one-hot encoding, we propose to treat style label as tokens and make multi-style combination with a more cleared representation for multi-style combination. We take the utilization of transformer to let style embedding interact with each of the visual embedding to enhance style impact. Basically, we solve the cross-domain style synthesis problem by jointly combine adversarial losses from different domain specified discriminators to supervise one caption. And we further design a simple but effective pseudo label self-training mixed with adversarial training to bring real training corpus (teacher forcing), thus improve the performance of style synchronized caption.

However, adversarial training in language generation (Yu et al. 2017) is harder than it in image generation, because there is no real gradient for discrete sampling process. In that case, we carefully design our adversarial loss into policy gradient formation in section . Based on our experiment result that while pushing generated sample distribution closer to biased style distribution, pure adversarial training without training captions may simultaneously be pushed deviated from real caption distribution. In that case, we propose to sample training corpus from real data and use discriminator to align masked style with pseudo label to real training corpus (teacher forcing) through certain threshold. At the third stage, we mixed adversarial and pseudo label training to enhance relevance and quality of synchronized multi-style captions. The whole training procedure is presented in Algorithm 1. The main contributions of this paper including three aspects:

- We propose a unique transformer-based model to solve cross-domain style synthesis for image caption (CDS-Cap), that is, to perform multiple styles in one generated sentence without synthetic multi-style datasets.
- We propose a multi-stage training framework to solve cross domain style synthesis problem for image caption by combining sequence adversarial learning and pseudo self-training. The proposed training frame work could learn different styles from single style datasets and align multiple styles in one caption.
- Experimental results show that our proposal is efficient in cross-domain style synthesis problems. Ablation study is reported to analyze the contribution of each stage.

Related Work

Image Captioning Most modern works in image caption presents an encoder-decoder (Mathews, Xie, and He 2016; Vinyals et al. 2015; Yang et al. 2017) formation. Here, encoders extract static global pooled (Vinyals et al. 2015) or regional pooled (Anderson et al. 2018) visual context from the input image. The decoders normally consist of basic structures for sequence modeling (Hochreiter and Schmidhuber 1997; Vaswani et al. 2017) which are in charge of receiving visual context from encoders and generate caption. Attention mechanism between encoder and decoder (Vinyals et al. 2015, 2017) has been proved of vital importance to generate high quality sentences. Recent trends including unsupervised image caption (Feng et al. 2019a), object or style controlled image caption (Guo et al. 2019; Zheng, Li, and Wang 2019), and using reinforcement learning to improve caption diversity (Rennie et al. 2017).

Stylized Image Caption Recent stylized image caption tasks could be categorized by training corpus: 1) paired caption with style (Mathews, Xie, and He 2016; Shuster et al. 2019) 2) unpaired caption by semi-supervised learning (Guo et al. 2019; Gan et al. 2017). SentiCap (Mathews, Xie, and He 2016) annotated training pairs with positive/negative styles and proposes to supervise stylized tokens generation with two parallel LSTM. StyleNet (Gan et al. 2017) annotated humorous/romantic styles and constructed input weight matrices with designed style factor. MSCAP (Guo et al. 2019) propose to train stylized language model with supervision of un-stylized caption reconstruction and style discriminator.

Cross-Domain Style transfer Most current methods in cross-domain style transfer or synthesis field could be summarized in two aspects. The first aspect utilizes the strong power of generative adversarial networks (Goodfellow 2017), where most approaches (Mirza and Osindero 2014; Zhu et al. 2017; Guo et al. 2019) force discriminator to learn to distinguish different styles form different domains then use the adversarial loss from discriminator to supervise generators. Especially, StarGAN (Choi et al. 2018) proves that cross-domain synthesis via GAN training mask and label mask is feasible. The second aspect is widely used in text generation area. Text style translation (Shen et al. 2017) proposes to encode different styles into a same latent space, and performs style transfer by using domain specified generator. Also, variational auto-encoders (VAE) (Larsen et al. 2016) could also perform style transfer by feeding conditional latent variables. In this paper, we mostly follow intuition introduced by StarGAN, and use adversarial loss to supervise generators.

Model Details

Task Definition

The cross-domain stylized image caption task could be formulated as follows: given K groups of training corpus $C = \{\{(\mathbf{I}_i, \mathbf{x}_i, s_i^k)\}_{i=1}^{N_k}\}_{k=1}^{K}$, where N^k is the number of Image-Caption pairs, $\mathbf{x}_i = \{x_i^1, x_i^2, \ldots, x_i^{T_i}\}$ is a caption sentence consisting of T_i words, s_i^k notes the associated style label from domain (dataset) k of the caption $\mathbf{x}_i$, the objective is to learn the conditional probability $P(\mathbf{x}|\mathbf{I}, \mathbf{s})$ from the corpus. Here, the attribute $\mathbf{s}$ is the combination of style labels $s^1, s^2, \ldots, s^k$ from k different domains(eg. Style label s^1 may come from sentiment domain, which consists of SentiCap dataset (Mathews, Xie, and He 2016) with value 'positive' or 'negative'. Meanwhile, style label s^2 may come

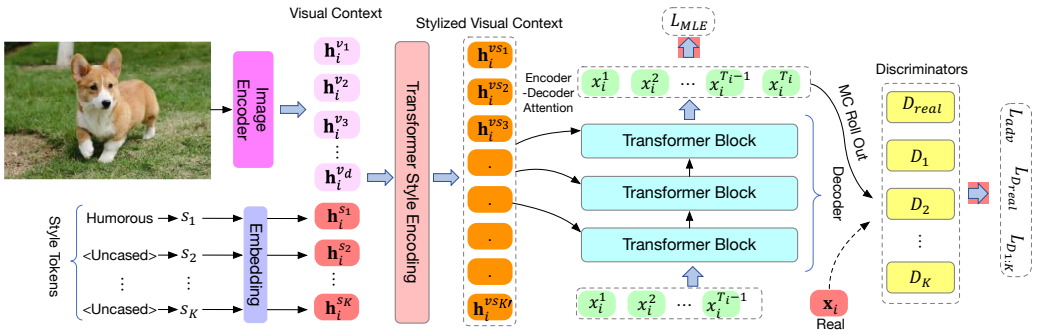

Figure 2: Caption decoder structure. Style embeddings are concatenated and interacted with each visual embedding.

from style domain, suggesting 'romantic' or 'humorous' style from Flickr10k (Gan et al. 2017) dataset). The model is proposed to learn from different domains and simultaneously control various styles.

Image Encoder

Given an image $\mathbf{I}$, the encoder uses deep convolutional neural networks (CNN) to extract semantic features from it. Here, existing methods could be roughly categorized into two aspects, 1) using a static, global pooled representation of the image (Vinyals et al. 2015) or 2) using a serious of spatial visual features (Anderson et al. 2018). Bottom-Up and Top-Down Attention (Anderson et al. 2018) has proved that, by proposing image regions (based on Faster R-CNN), each with an associated feature vector and incorporates these features through attention, it could reach sate-of-the-art performance on image caption task (Anderson et al. 2018). Since the main focus of this paper is cross-domain style synthesis, we direct uses the the same structure proposed in Bottom-up and Top-Down Attention (Anderson et al. 2018), where it uses Faster-RCNN (Ren et al. 2017) with base-net ResNet101 (He et al. 2016) as the image feature extractor. After performed non-maximum suppression, the output of Faster-RCNN is a set of regions and features. Given region j, we flatten regional pooled features as one visual context embedding $\mathbf{h}_{v_j}$ for transformer decoders as it shows in Fig. 2.

Caption Decoder

We design a style-conditional caption decoder $G(\cdot)$ that captures the language property distributions of potential style combinations. The caption decoder is proposed to satisfy two aspects of requirements, 1) it should be able to generate tokens that follow the language modeling 2) it should be able to adjust biased token generation probabilities according to given style combinations $\mathbf{s} = s^1, s^2, \ldots, s^k$, where s^k denotes style label from paired annotations from domain k. Instead of using LSTM as decoder as most previous papers (Guo et al. 2019; Feng et al. 2019b), this paper proposes a caption decoder with single directional transformer (Vaswani et al. 2017). In this subsection, detailed model structure of the caption decoder $G(\cdot)$, and style combination input methods are respectively described below.

Style Labels as Tokens With Masks Different from paper (Guo et al. 2019) which mixes different domains and uses a one-hot vector with shape $k+1$ to represent k styles, we propose to use K tokens $\{s^k\}_{k=1}^K$ to represent styles. Here, each token s^k could have j^k values(eg. $s^k \in \{Humorous, Romantic\}, j^k = 2$). These style tokens $\{s^k\}_{k=1}^K$ are projected to style embeddings $\{\mathbf{h}^{s_k}\}_{k=1}^K$ through standard word2vec as other word tokens. We propose this formation to realize a more clear style representation combination from multiple different domains. However, after careful observations of the caption datasets, we conclude that some of the caption are uncased and not belongs to any categories. In that case, we use token $< Uncased >$ as a style mask in each s^k. This setting also benefits the first stage conditional pre-training described in section . The style embedding vectors $\{\mathbf{h}^{s_k}\}_{k=1}^K$ and concatenate as extra length K along with calculated visual context $\{\mathbf{h}^{v_j}\}_{j=1}^d$ by image encoder defined in section .

Caption Generator Structure with Transformers. Different from pure language modeling such as original GPT-2 (Radford et al. 2019), we modify our caption generator similarly to neural response generation systems. Given style input embedding $\{s^k\}_{k=1}^K$ and visual context $\{\mathbf{h}^{v_j}\}_{j=1}^d$, the caption generator is proposed to generate caption sequences $\mathbf{x} = x^1, x^2, ..., x^{T_i}$. The decoding process of token x^{T_i} could be defined as:

$$G(x^{T_i}|\{\mathbf{h}^{v_j}\}_{j=1}^d, \{s^k\}_{k=1}^K, x^1, x^2, ..., x^{T_i-1}) \quad (1)$$

The detailed structure of the proposed caption decoder model is shown in Fig. 2. The style tokens from K domains are concatenated and embedded through an embedding layer. Then the style embedding $\{\mathbf{h}^{s_k}\}_{k=1}^K$ are concatenated with visual context $\{\mathbf{h}^{v_j}\}_{j=1}^d$ and are fed to a bidirectional transformer block to obtain stylized visual context embedding $\mathbf{h}^{vs}$ as shown in Fig. 2. The propose of this block is to make each of the style embeddings to interact with each of the visual embedding thus enlarge the impact of style feature. The single directional transformer decoder receives stylized visual context through encoder-decoder attention as

defined in Eq. 2.

$$\begin{aligned} \mathbf{h}^{dec_j} &= MH(\mathbf{h}^{vs}, \mathbf{h}^{dec_{j-1}}, \mathbf{h}^{vs}) \\ \mathbf{h}^{dec_0} &= MH(\mathbf{k}, \mathbf{q}, \mathbf{v}), \\ \mathbf{k} = \mathbf{v} &= [x^1, x^2, \ldots, x^{T_i - 1}], \mathbf{q} = \mathbf{h}^{vs} \end{aligned} \tag{2}$$

where $MH(\cdot)$ denotes multi-head attention block introduced in transformer (Vaswani et al. 2017), and $\mathbf{k}, \mathbf{q}, \mathbf{v}$ respectively denotes input of query projection, key projection and value projection for MH block. Besides the first decoding block, query $\mathbf{q}$ of each decoding transformer is a stylized visual feature $\mathbf{h}^{vs}$. Then the current token x^{T_i} is sampled through equation:

$$x^{T_i} \sim p^{T_i} = softmax(\mathbf{W}\mathbf{h}^{dec_j}), \tag{3}$$

where $\mathbf{W}$ denotes parameters project hidden states to ouput size and x^{T_i} is sampled according to probability distribution p^{T_i}.

Cross-Domain Style Synthesis

As mentioned in section that one major challenge of cross-domain caption style synthesis is that, for each training caption $\mathbf{x}$ corpus from domain k, only style label s^k is given. To solve this challenge, we propose a novel multi-stage training framework for cross-domain style synthesis intuitively shown in Fig. 3. The training framework could be presented sequentially in three stages as below: 1) pre-training conditional caption generator on each domain with style masks (section) 2) adversarial training each generated caption with designed multi-domain adversarial loss (section) 3) pseudo label self-training to enhance synchronized generation quality (section). Moreover, we give an overall training procedure in Algorithm 1 for a clear representation.

Pre-Training with Masked Style Combination

In this stage, we propose to firstly establish basic associations between input image, style labels and output caption sequences. Before starting this stage, the image extractor has been separately trained on COCO (Lin et al. 2014) and visual Genome (Krishna et al. 2017) follow (Anderson et al. 2018) and fixed. However, the generator defined in section requires style label from every domain to construct the style feature embedding. In order to solve addressed problem that each caption only has style label from domain k or does not have any style labels, we mask unknown style labels by mask token $< Uncased >$. (For example, given only label s^2 as $Humorous$, the style combination input is constructed as $\{s^1, s^2, s^3\} = \{< Uncased >, Humorous, < Uncased >\}$.) It is noted that, in training corpus from domain k, although a certain caption is annotated as style A, it may or may not contains intrinsic style B but not yet be labeled explicitly. Consider giving style label s^k from domain k, we propose $< Uncased >$ token mask to represent the randomness of style distributions from other domains. This assumption is rational due to the compared large caption training corpus scale. In this case, the whole system except fixed image feature extractor is trained through maximum likelihood estimation loss (MLE loss) as defined below:

$$\begin{aligned} L_{MLE} = -E_{x^{T_i}} \log(G(x^{T_i} | \{\mathbf{h}^{v_j}\}_{j=1}^{d}, \{s^k\}_{k=1}^{K}, \mathbf{X})) \\ x^{T_i} \sim p_{real}, \mathbf{X} = x^1, x^2, ..., x^{T_i - 1} \end{aligned} \tag{4}$$

After training in this stage, conditional caption generator $G(\cdot)$ is expected to 1) obtain basic abilities to generate caption with correct semantic meaning according to given visual context 2) establish compared loose relationship between given style and corresponding style in caption (because the training captions have already contains certain styles). However, style tokens from different domains never co-exist in this stage. Thus, the style constraints are still rough and imprecise as it shows in experiment. We propose to solve the style co-exist problem by joint adversarial loss from multiple domains specified discriminators as described below.

Adversarial Training for Style Synthesis

Remind that, after the first stage training of masked style combinations from different domains, the model is expected to learn a rough stylized distribution. However, the caption (language) generator has not been trained with style combination labels which contains various styles at the same time. Meanwhile, research in text style analysis area (Maas et al. 2011; Dathathri et al. 2019) suggest that transformers with moderate parameters have a good ability to distinguish different text styles. We propose to solve cross-domain style synthesis by **synchronously** using the style adversarial discriminators from K **different domains** to supervise one proposed conditional caption generator. The details of how to obtain discriminator and adversarial train the caption generator have been presented below.

Domain Specified Style Discriminator Instead of presenting one discriminator, we propose $K + 1$ different discriminators, which consist of one real/fake discriminator and K domain style discriminators to distinguish $j_k + 1$ different styles (as mentioned in section , j_k styles and one $< Uncased >$ mask). After pre-training at the first stage and before generative adversarial training at the second stage, we pre-train discriminators in their domain. The real/fake discriminator is trained by loss:

$$\begin{aligned} L_{D_{real}} = -E_{\mathbf{x}} \log D_{real}(\mathbf{x}) \\ -E_{\mathbf{I},\mathbf{s}} \log(1 - D_{real}(G(\mathbf{I}, \mathbf{s}))), \\ \mathbf{x} \sim p_{real}, \mathbf{I} \sim p_{real}, \mathbf{s} \sim Rand, \end{aligned} \tag{5}$$

where $D_{real}(\mathbf{x})$ is a probability indicating how likely a caption $\mathbf{x}$ is from real sequence data or not. $D_{real}(\mathbf{x})$ is normally a small classification network in the area of adversarial learning (Goodfellow et al. 2014). Real captions are sampled from training corpus and fake captions are generated by pre-trained caption generator with real image samples and a randomly sampled style as inputs. Domain specific discriminators are trained separately in different domains by cross-entropy with real training corpus from K domains and unstylized domain. The training loss of discriminator k is defined as:

$$L_{D_k} = -E_{s^k \sim p_{real}} \log(D_k(s^k | \mathbf{x})), \tag{6}$$

where style discriminators are expected to predict right style distributions given caption $\mathbf{x}$.

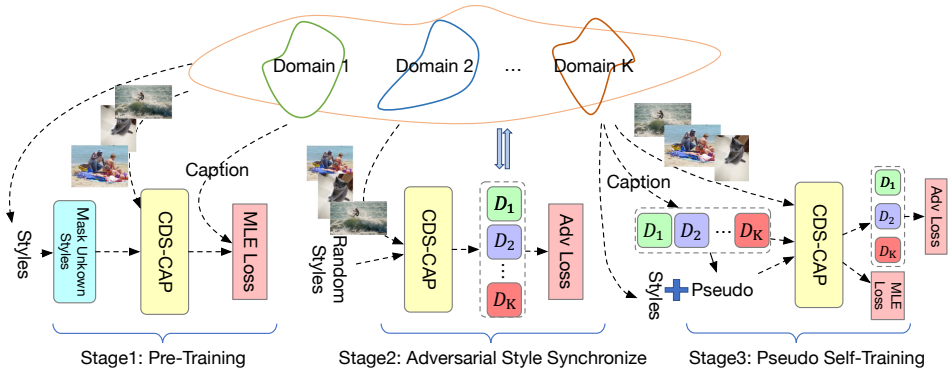

Figure 3: Multi-stage training flow. Here, pre-training and pseudo self-training uses training corpus with image, style and captions, while adversarial training only use images from real training corpus.

Generative Adversarial Loss with Roll Out Generative Adversarial networks (Choi et al. 2018; Goodfellow 2017) are efficient approaches for cross-domain style transfer. However, discriminator can't acquire gradient through caption generation, which consists a discrete decoding (sample) process. To solve this problem we use policy gradients in a reinforcement learning formation as mentioned in sequence GAN (Yu et al. 2017; Wang and Wan 2018). Different from (Wang and Wan 2018) which simply use separate discriminator loss for different generators, we combine losses from K domains to supervise the generator simultaneously to force caption generator to synchronize styles from different domains. The adversarial loss is present as reward in policy gradient as defined in Eq. 7.

$$L_{adv} = G(x^{T_i}|\mathbf{s}, \mathbf{I}, x^{1:T_i-1})R(x^{1:T_i}, \mathbf{s}), \tag{7}$$

where $\mathbf{s}$ denotes style combinations $\{s^k\}_{k=1}^K$, $G(x^{T_i}|\mathbf{s}, \mathbf{I}, x^{1:T_i-1})$ is the simplified representation for generative model in Eq. 1, and $R(x^{1:T_i}, \mathbf{s})$ denotes reward function given current token and style $\mathbf{s}$. Because there is no sense to let discriminators only judge on a incomplete sequence, Monte Carlo search (MC search) is applied to sample and average unknown $|\mathbf{x}| - T_i$ tokens(where $|\mathbf{x}|$ is the total length).

$$R(x^{1:T_i}, \mathbf{s}) = \begin{cases} \frac{1}{N}\sum_{n=1}^{N} \lambda_{real} D_{real}(x_n^{1:|\mathbf{x}|}) \\ + \sum_{k=1}^{K} \lambda_k s^k \log D_k(\tilde{s}_k | x_n^{1:|\mathbf{x}|}), T_i < |\mathbf{x}| \\ \lambda_{real} D_{real}(x_n^{1:|\mathbf{x}|}) \\ + \sum_{k=1}^{K} \lambda_k s^k \log D_k(\tilde{s}^k | x^{1:T_i}), T_i = |\mathbf{x}|, \end{cases} \tag{8}$$

where $x_n^{1:|\mathbf{x}|}$ denotes n^{th} sampled sequence by MC search, λ_k is the coefficient parameter according to cross-entropy loss between given style $\mathbf{s}$ and predicted style $\tilde{s}^k$(by pre-trained generator). The final reward is calculated by average rewards from N times MC search for policy gradients. It should be noted that, during adversarial generative search, we randomly sample different style combinations $\mathbf{s}$, which allows styles from different domains at the same time. The well-trained discriminator provides a reward for these generated captions with synchronized styles. It is noted that, in this stage, no training corpus is used in the training caption generator. The discriminator and caption generator is alternately trained step by step as described in Algorithm 1.

Self-Training with Pseudo Label

Mode collapse has been a serious problem especially for text adversarial training (Yu et al. 2017). Since at the second stage, adversarial training is merely between generated samples and adversarial losses, the lack of real samples may lead to the generator prefer to generate sample to gain high reward from discriminator. However, it is observed from our experiment in section that, pure adversarial training may push generated sample deviated from real sample distribution. In that case, in third stage, we propose to perform a self-training formation. We sample real training corpus from certain domain and use domain specified discriminators to judge the confidence of presenting styles from other domains. Once the predicted score is larger than a certain threshold (eg. 0.8) we assign masked styles with predicted pseudo label. Given pseudo label, we could train the generator alternately with both proposed MLE loss in Eq. 4. and adversarial losses. Intuitively, we could constrain generated closer to real distribution while maintaining different styles by alternately present pseudo self-training and adversarial training in the third stage.

Given training corpus $\{\mathbf{I}, \mathbf{x}, \mathbf{s}\}$, where $\mathbf{s}$ should only contain one style label with other style masked (eg. $\mathbf{s} = \{< Uncased >, Humorous, < Uncased >\}$), we align pseudo label to masked style tokens through discriminator selection as described above. The well-trained discriminators evaluate confidence that one real sample from certain domain simultaneously presenting styles from other domains. And the pseudo label is only aligned to mask token only if the predicted score has reach threshold (eg. we set the threshold as 0.8). For example, if $D_1(positive|\mathbf{x}) > 0.8$, we directly assign style $\mathbf{s}$ as $\{Positive, Humorous, < Uncased >\}$. We search training corpus that successfully assign pseudo labels, and perform self-training on these pseudo training corpus with MLE loss L_{MLE} defined in Eq. 4. At the third stage, the pseudo label self-training is performed alternately with adversarial training defined in the second stage. Experiment results show that, by pseudo label

Algorithm 1 Multi-stage training procedure for cross-domain style synthesis

Require: Training corpus $C = \{\{(\mathbf{I}_i, \mathbf{x}_i, s_i^k)\}_{i=1}^{N_k}\}_{k=1}^{K}$, Generator G, Discriminator D_{real}, Style Discriminator $\{D_k\}_{k=1}^{K}$ for K domains with parameter θ_G, $\theta_{D_{real}}$,$\{\theta_{D_k}\}_{k=1}^{K}$.

// Stage-1: Pre-training with masked label

Initialize the model parameters θ_G, $\theta_{D_{real}}$, $\{\theta_{D_k}\}_{k=1}^{K}$

for Step $<$ total step, sample $(\mathbf{I}_i, \mathbf{x}_i, s_i^k) \sim C$ **do**

  Mask unknown domain style, $\mathbf{s} = mask(s_i^k)$

  Train θ_G by minimizing L_{MLE} (Eq. 4)

// Stage-2: Adversarial Training

for domain $k < K$ **do** ▷ domain style classifier

  Train θ_{D_k} by minimizing L_{D_k} (Eq. 6)

for Step $<$ total step, sample $(\mathbf{I}_i, \mathbf{x}_i) \sim C$ **do**

  Generate fake caption $(\tilde{\mathbf{x}}_i) \sim G(\mathbf{I}, \mathbf{s})$, with random style combination $\mathbf{s}$

  Train $\theta_{D_{real}}$ by minimizing $L_{D_{real}}$ (Eq. 5)

while Step $<$ adversarial step **do**

  for g_{step} **do** ▷ update generator

    Train θ_G by minimizing L_{adv} (Eq. 7)

  for d_{step} **do** ▷ update discriminator

    Train $\theta_{D_{real}}, \theta_{D_{1:K}}$ by minimizing $L_{D_{real}}$ and $L_{D_{1:K}}$

// Stage-3: Pseudo Label Fine-tuning

for $pseudo_{step}$ **do** ▷ self-training with real data

  Train θ_G by minimizing L_{MLE}

  Execute adversarial training in stage 2 for one step.

self-training, we could further make the synchronized captions closer to the real data distribution.

Training Procedure

Each stage of the training procedure has been provide details and the training target as above. In this subsection, we provide a clear overall pip-line using an algorithm form as it shows in Algorithm 1. To summarize, at the first we propose to let the caption generator learn essential abilities to generate correct caption sentences according to given visual context by L_{MLE}. Also at this stage, the caption generator should have learn rough and imprecise style representation, as at this stage, the model is trained with masked single style captions. At the second stage, we propose to jointly use discriminators from K domains to supervise the caption generation process simultaneously, which also forces the generator to synchronize styles from different domains in one caption. It is noted that no real samples are used in this stage. In order to avoid mode collapse[1] and generative deviation from real distribution, we propose pseudo label self-training on selected training corpus as an addition to adversarial training at stage three. Different from stage two which only learn from style discriminators (style loss), we directly select real samples which performs multi-styles simultaneously by using trained style discriminators, and assign pseudo labels for these real samples as training data. After the proposed three-stage training procedure, our proposed model should have the ability to control caption styles from different domains at the same time.

[1]Normally denotes to models only generate meaningless content which only make sense to discriminators.

Experiment

In this section, we respectively introduce experiment datasets in section , and implementation details in section . Automatic evaluation metrics and ablation study analysis are reported in section . We show the caption cases in section . Experiment results suggest that our proposal is efficient to generate captions with cross-domain styles.

Dataset

The experiments is implemented on two public stylized caption datasets, where FlickrStyle 10K (Gan et al. 2017) contains 14K captions with style 'Humorous' and 'romantic', and SentiCap (Mathews, Xie, and He 2016) contains 8922 captions with style 'Positive' and 'Negative'. Original COCO caption (Lin et al. 2014) dataset is used for pre-training and provide sample with mask label $< Uncased >$ as presented in section . This experiment split train and test sample follow (Lin et al. 2014), where training, validation and test set contains 82783, 40504 and 40775 images respectively. For the two stylized datasets, we randomly split 5% for test and 5% for validation. For representing convenience, we denote the 'Humorous', 'Romantic', 'Positive', 'Negative' styles, and uncased style (also present as un-stylized mask in section respectively as H, R, P, N, and C. Only training split is used for training procedures.

Implementation Details

For image feature extractor, we keep parameter settings and pre-training methods of Faster-RCNN the same with Bottom Up, Top Down paper (Anderson et al. 2018) except for the output layer (**CDS-**). We also report results by using static CNN as feature extractor to keep it the same with previous stylized caption papers (**CDS-***). The output of each sliced global pooled visual feature embedding and the regional visual feature embedding share the same embedding dimension 1024. The visual feature's length is set as 32, which consists of 16 global features and 16 regional visual features each with embedding dimension 1024. Accordingly, the embedding size of position embedding and segmentation embedding is the same as 1024. These two embeddings and the stylized token embeddings are respectively generated by three separate word2vec layers with trainable parameters. The hidden size of all transformer blocks in our model is 512. We stack two bidirectional transformer blocks as the BERT Encoder and three single directional transformer blocks as the decoder. For caption decoder word embedding, we set the embedding size as 200. We use 2 simple bidirectional transformer blocks with hidden size 256 for each of the discriminator. The word embedding of all discriminators are shared with decoder blocks.

All parameters not specially mentioned are initialized with Xavier (Glorot and Bengio 2010) method. We use Adam (Kingma and Ba 2015) as the training optimizer. For pre-training we use learning rate 5×10^{-4} and for other training processes, we use learning rate 1×10^{-4}. As mentioned in section , the experiments are applied on two domains (SentiCOCO, and FlickrStyle 10K), where uncased

Table 1: Metrics on stylized data. CDS-* denotes the same model with different style inputs, where PU, NU, UH, UR respectively denotes single style positive, negative, humorous and romantic. Here U denotes the uncased mask. We also give evaluation results of force generator to generate unseen style combinations such as PH, PR to support efficiency of our multi-style control. The footnote ^{s} denotes our model with the same static CNN feature extractor with previous methods.

Positive						Negative					
Method	B-1	B-3	M	C	cls.	Method	B-1	B-4	M	C	cls.
StyleNet	46.7	13.1	13.2	56.0	67.3	StyleNet	47.1	13.4	14.0	57.1	66.7
SentiCap	50.5	20.1	16.7	63.0	77.4	SentiCap	51.3	20.6	17.5	64.1	76.3
MSCAP	47.1	16.8	17.1	55.0	92.5	MSCAP	45.9	16.3	16.9	55.6	92.8
CDS-PU^s	59.5	23.6	21.3	72.1	95.3	**CDS-NU**^{s}	59.7	26.3	20.2	69.4	95.7
CDS-PH^s	55.9	18.4	19.0	65.2	93.7	**CDS-NH**^{s}	52.6	16.8	17.9	64.3	90.2
CDS-PR^s	54.3	17.8	18.4	64.2	92.2	**CDS-NR**^{s}	53.6	17.3	16.6	62.9	89.7
CDS-PU	61.3	26.7	22.8	74.2	96.8	**CDS-NU**	61.8	27.1	23.1	73.7	95.6
CDS-PH	56.7	18.9	19.6	67.3	93.7	**CDS-NH**	53.4	17.8	18.1	65.4	90.1
CDS-PR	55.3	18.1	19.3	66.1	93.0	**CDS-NR**	54.1	17.7	17.4	63.4	89.2
Humorous						Romantic					
StyleNet	13.7	1.9	4.9	7.8	43.1	StyleNet	13.2	2.0	5.1	8.0	44.9
SentiCap	29.2	9.1	12.1	38.2	82.4	SentiCap	27.8	8.8	11.5	36.9	84.0
MSCAP	17.2	2.3	5.6	11.0	88.9	MSCAP	16.8	1.7	5.5	37.1	86.5
CDS-UH^s	34.3	9.0	12.9	43.9	91.2	**CDS-UR**^{s}	38.5	10.3	12.7	42.8	90.2
CDS-PH^s	29.6	8.9	12.2	38.4	89.1	**CDS-PR**^{s}	30.4	9.8	12.4	41.5	90.3
CDS-NH^s	27.1	8.0	9.6	35.2	84.5	**CDS-NR**^{s}	26.9	7.2	9.2	36.0	81.8
CDS-UH	37.3	10.3	13.5	44.7	90.7	**CDS-UR**	39.3	10.8	13.9	43.2	91.0
CDS-PH	31.3	9.7	12.7	39.5	89.3	**CDS-PR**	30.9	10.5	13.2	42.3	90.0
CDS-NH	28.4	8.3	10.1	36.9	84.2	**CDS-NR**	27.8	7.9	9.8	36.8	82.5

(un-stylized) sample is from normal coco dataset. Training coefficient λ_{real} for fake/true discriminator D_{real}, and λ_1, λ_2 for style discriminator D_1, D_2 are respectively set to 0.5, 0.15 and 0.15. For stage 1 pre-training with masked label , we train the model as normal image caption model for 15 epochs with learning rate 5×10^{-4}. For stage 2 adversarial training, we set the maximum training step as 25000, where the ratio between g_{step} and d_{step} is 5:1 (Arjovsky, Chintala, and Bottou 2017) with learning rate 1×10^{-5}. For stage 3 pseudo fine-tuning , we set the maximum training step as 15000 with learning rate 1×10^{-5}.

Compared Approaches

Only a few works address stylized caption generation tasks. Most of them only consider single style generation including **SentiCap** (Mathews, Xie, and He 2016) and **StyleNet** (Gan et al. 2017). **MSCAP** (Guo et al. 2019) consider unpaired multiple style generation under similar dataset setting. However, all of the previous papers treat styles from different domain exclusive to each other and ignore the cross-domain style synthesis problem. We select MSCAP to compare the performance of a single style generation. In order to illustrate the efficiency of our proposed method, we directly compare our multi-style generation method with mentioned single-style generation methods. By comparing with these approaches, we illustrate that our proposed cross-domain synthesis could let the caption generator generate captions with multiple styles at the same time, while keeping the generation of each single style with good quality.

Evaluation Metrics

In this section, we use automatic evaluation metrics to compare our proposal with previous methods.

Generative Rationality It is noted that, the generated captions should firstly be meaningful and fluent as a sentence itself before considering the stylized properties. In order to illustrate the stylized captions are meaningful, we directly use semantic relevance measure metrics used in general image caption methods (Anderson et al. 2018; Lavie and Agarwal 2007). We use the widely used automatic evaluation metrics, BLEU (Papineni et al. 2002), METEOR (Lavie and Agarwal 2007), ROUGE-L (Lin 2004), CIDEr (Vedantam, Zitnick, and Parikh 2015) and SPICE (Anderson et al. 2016) to measure generative relevance from real distribution of our proposed model.

Multi-Stylized Measurement Since we haven't found papers about presenting multiple image caption styles at the same time, we compare our proposal with single style caption approaches. We use our proposed method to generate captions with multiple style combinations, and firstly test the rationality evaluation metrics mentioned above and shows it in Tab. 1. Simultaneously, to evaluate styles, we separated train a style classifier for each style. We use the style classification accuracy to illustrate caption style property. The automatic evaluation metrics of 'Positive' and 'Negative' are calculated on SentiCOCO dataset, and 'Humorous' and 'Romantic' are calculated on FlickrStyle 10K dataset.

It is noted that, for more fair comparison, we also report metrics by using the same feature extractor (static CNN) as

Table 2: Generated sample of multi-style control

Sample	Humorous & Positive	Romantic & Positive	Humorous & Negative	Romantic & Negative	Uncased
	a nice polar bear with something white is playing with water	a lovely polar bear is swimming in the water, call it puppy love	a stupid polar bear is playing with dirty water	a dirty polar bear is swimming in the loving water	a polar bear is swimming in the water
	a funny man riding a wave on the top of nice board	a funny man riding on his loving surfboard in the water	a scared man riding on a wave on surfboard	a lovely man on a dirty wave with his loving surfboard	a man riding a wave on the top of the surfboard

Table 3: Influence bring by different training stages

Method	Train with Style	BLUE-1	BLUE-3	METEOR	ROUGE-L	CIDEr	SPICE
CDS-Base	No	78.9	46.9	27.1	57.1	115.8	21.2
CDS-Mask	Yes	75.7	43.6	24.9	52.1	98.8	18.2
CDS-Adv	Yes	63.2	36.9	21.8	48.3	80.2	16.9
CDS-Pseudo	Yes	74.2	42.6	23.6	51.0	83.1	17.1

previous methods(MSCAP, StyleNet). It could be observed that our model generally perform better than previous methods. Also, our model achieves the best performance while using detector based feature extractor. We argue this result is rational based on two points: 1) our model use transformer as a generator, this would bring stronger language model and model capacity. 2) MSCAP and StyleNet use unpaired data for training. However for our model, besides the unpaired adversarial training, we select real samples by style discriminator and train paired data with pseudo label. It is also noted that for style 'Negative & Humorous' and 'Negative & Romantic', metric scores are significantly lower than others. We argue this is also rational because tokens with humorous and romantic styles are normally do not have strongly related distribution jointly with negative.

Generated Samples Generated samples of different style combinations have been reported for intuitive understanding in Tab. 2. It could be observed that instead of only presenting one style, our model could simultaneously control styles from different domains. Please note that not all of the forced multi-style generation could reach the same perfect quality. It is rational because some of the given image is hard to present styles from both domains at the same time. The lack of training data limited language model to learn enough joint distributions with multiple styles.

Ablation Study

We conduct our ablation study at two aspects: 1) how the stylized training influence the original caption model 2) how each component of the proposed framework influences the style classification accuracy.

Comparison with Un-stylized Caption We first test the un-stylized evaluation metrics on the standard COCO dataset of CDS model after each training stage and report our result in Tab. 3. We show the evaluation metrics on COCO dataset by setting style s with all masks (Uncased) to analyze how stylized training would influence the original ability for image caption. It is observed that pure adversarial training without captions from real training corpus (only CDS-Adv does not include real captions for training) will lower the evaluation scores. This observation supports our intuition that pure adversarial training may lead to generator prefers to generate samples which could acquire higher reward to fool discriminators, but it may simultaneously push generator distribution deviated from real caption distribution. (Un-stylized caption CDS-Base acquires the highest evaluation sore.) And by adding pseudo label training with real data, we force the generated caption closer to the original distribution.

Table 4: Ablation study of style accuracy change through multi-stage training

	Positive			Humorous		
	B-1	B-3	cls.	B-1	B-3	cls.
CDS-Mask-PU	62.1	27.3	91.8	-	-	-
CDS- Mask-UH	-	-	-	38.1	11.9	86.7
CDS-Mask-PH (without synthesis)	40.7	10.3	77.6	16.1	0.7	70.4
CDS-Adv-PH (start synthesis)	47.1	17.2	92.0	20.2	1.6	93.4
CDS-Pseudo-PH (pseudo fine-tune)	56.7	18.9	93.7	31.3	9.7	89.3

Contribution from Each Attribute of the Model In order to figure out the contributions of each component to our final cross-domain style synthesis performance, we measure how styles accuracy and relevancy change throughout the training process. Taking $Positive\&Humorous$ as an example, we show how styles accuracy and BLEU-1,BLEU-3 score changes on SentiCOCO dataset and shows in Tab. 4. It could be observed that, while the caption model is only trained by style masks, forcing caption generator to simultaneously present multiple styles only could acquire poor evaluation metrics. Experimental results suggest that

adversarial training largely increases style classification accuracy, which illustrates the efficiency of our proposed adversarial synthesis. As mentioned above, adversarial training may harm relevancy (reflected by BLUE) between generated sample and real captions. Observed metrics support that pseudo self-training mixed with adversarial training successfully increase performance on BLEU scores.

Conclusion

In this paper, we present a novel transformer-based image caption model and accordingly the multi-stage training framework to solve cross-domain caption style synthesis problem. Experiment results show the efficiency of our proposed methods. Forcing the image caption model to simultaneously present multiple styles cross-domain is still a hard question. We will explore how to generate more detailed or object-oriented captions in the future work. Also, in this paper, we only use a simple position embedding for transformers. However, we think how to properly present spatial relations to the language model is a promising task, too.

References

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. SPICE: Semantic Propositional Image Caption Evaluation. In Leibe, B.; Matas, J.; Sebe, N.; and Welling, M., eds., *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*, volume 9909 of *Lecture Notes in Computer Science*, 382–398. Springer. doi:10.1007/978-3-319-46454-1_24.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 6077–6086. IEEE Computer Society. doi:10.1109/CVPR.2018.00636.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein Generative Adversarial Networks. In *International Conference on Machine Learning*, 214–223.

Chen, C.; Mu, S.; Xiao, W.; Ye, Z.; Wu, L.; and Ju, Q. 2019. Improving Image Captioning with Conditional Generative Adversarial Nets. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, 8142–8150. AAAI Press.

Choi, Y.; Choi, M.; Kim, M.; Ha, J.; Kim, S.; and Choo, J. 2018. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 8789–8797. IEEE Computer Society.

Dathathri, S.; Madotto, A.; Lan, J.; Hung, J.; Frank, E.; Molino, P.; Yosinski, J.; and Liu, R. 2019. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. *CoRR* abs/1912.02164.

Donahue, J.; Hendricks, L. A.; Rohrbach, M.; Venugopalan, S.; Guadarrama, S.; Saenko, K.; and Darrell, T. 2017. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. *IEEE Trans. Pattern Anal. Mach. Intell.* 39(4): 677–691.

Feng, Y.; Ma, L.; Liu, W.; and Luo, J. 2019a. Unsupervised Image Captioning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 4125–4134. Computer Vision Foundation / IEEE. doi:10.1109/CVPR.2019.00425.

Feng, Y.; Ma, L.; Liu, W.; and Luo, J. 2019b. Unsupervised image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4125–4134.

Gan, C.; Gan, Z.; He, X.; Gao, J.; and Deng, L. 2017. StyleNet: Generating Attractive Visual Captions with Styles. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 955–964. IEEE Computer Society. doi:10.1109/CVPR.2017.108.

Glorot, X.; and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In Teh, Y. W.; and Titterington, D. M., eds., *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, volume 9 of *JMLR Proceedings*, 249–256. JMLR.org.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

Goodfellow, I. J. 2017. NIPS 2016 Tutorial: Generative Adversarial Networks. *CoRR* abs/1701.00160.

Guo, L.; Liu, J.; Yao, P.; Li, J.; and Lu, H. 2019. MSCap: Multi-style image captioning with unpaired stylized text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4204–4213.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 770–778. IEEE Computer Society.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8): 1735–1780.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.; Shamma, D. A.; Bernstein, M. S.; and Fei-Fei, L. 2017. Visual

Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision* 123(1): 32–73.

Lake, B. M.; Ullman, T. D.; Tenenbaum, J. B.; and Gershman, S. J. 2017. Building machines that learn and think like people. *Behavioral and brain sciences* 40.

Larsen, A. B. L.; Sønderby, S. K.; Larochelle, H.; and Winther, O. 2016. Autoencoding beyond pixels using a learned similarity metric. In Balcan, M.; and Weinberger, K. Q., eds., *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, 1558–1566. JMLR.org.

Lavie, A.; and Agarwal, A. 2007. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In Callison-Burch, C.; Koehn, P.; Fordyce, C. S.; and Monz, C., eds., *Proceedings of the Second Workshop on Statistical Machine Translation, WMT@ACL 2007, Prague, Czech Republic, June 23, 2007*, 228–231. Association for Computational Linguistics.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W04-1013.

Lin, T.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In Fleet, D. J.; Pajdla, T.; Schiele, B.; and Tuytelaars, T., eds., *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, 740–755. Springer.

Maas, A. L.; Daly, R. E.; Pham, P. T.; Huang, D.; Ng, A. Y.; and Potts, C. 2011. Learning Word Vectors for Sentiment Analysis. In Lin, D.; Matsumoto, Y.; and Mihalcea, R., eds., *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, 142–150. The Association for Computer Linguistics.

Mathews, A. P.; Xie, L.; and He, X. 2016. SentiCap: Generating Image Descriptions with Sentiments. In Schuurmans, D.; and Wellman, M. P., eds., *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, 3574–3580. AAAI Press.

Mirza, M.; and Osindero, S. 2014. Conditional Generative Adversarial Nets. *CoRR* abs/1411.1784.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, 311–318. ACL.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1(8): 9.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39(6): 1137–1149.

Rennie, S. J.; Marcheret, E.; Mroueh, Y.; Ross, J.; and Goel, V. 2017. Self-Critical Sequence Training for Image Captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 1179–1195. IEEE Computer Society. doi:10.1109/CVPR.2017.131.

Shen, T.; Lei, T.; Barzilay, R.; and Jaakkola, T. S. 2017. Style Transfer from Non-Parallel Text by Cross-Alignment. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, 6830–6841.

Shuster, K.; Humeau, S.; Hu, H.; Bordes, A.; and Weston, J. 2019. Engaging Image Captioning via Personality. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 12516–12526. Computer Vision Foundation / IEEE.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, 5998–6008.

Vedantam, R.; Zitnick, C. L.; and Parikh, D. 2015. CIDEr: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 4566–4575. IEEE Computer Society.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 3156–3164. IEEE Computer Society. doi:10.1109/CVPR.2015.7298935.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2017. Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge. *IEEE Trans. Pattern Anal. Mach. Intell.* 39(4): 652–663.

Wang, K.; and Wan, X. 2018. SentiGAN: Generating Sentimental Texts via Mixture Adversarial Networks. In Lang, J., ed., *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, 4446–4452. ijcai.org. doi:10.24963/ijcai.2018/618.

Wen, T.; Gasic, M.; Mrksic, N.; Su, P.; Vandyke, D.; and Young, S. J. 2015. Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems. In Màrquez, L.; Callison-Burch, C.; Su, J.; Pighin,

D.; and Marton, Y., eds., *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, 1711–1721. The Association for Computational Linguistics.

Yang, L.; Tang, K. D.; Yang, J.; and Li, L. 2017. Dense Captioning with Joint Inference and Visual Context. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 1978–1987. IEEE Computer Society.

Yu, L.; Zhang, W.; Wang, J.; and Yu, Y. 2017. SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient. In Singh, S. P.; and Markovitch, S., eds., *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, 2852–2858. AAAI Press.

Zheng, Y.; Li, Y.; and Wang, S. 2019. Intention Oriented Image Captions With Guiding Objects. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 8395–8404. Computer Vision Foundation / IEEE. doi:10.1109/CVPR.2019.00859.

Zhou, L.; Xu, C.; Koch, P. A.; and Corso, J. J. 2016. Image Caption Generation with Text-Conditional Semantic Attention. *CoRR* abs/1606.04621.

Zhu, J.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2242–2251. IEEE Computer Society.