Unsupervised Stylish Image Description Generation via Domain Layer Norm

Cheng-Kuan Chen,[1*] **Zhufeng Pan,**[1*] **Ming-Yu Liu,**[2] **Min Sun**[1]
[1]Department of Electrical Engineering, National Tsing Hua University [2]NVIDIA
{sttagomantis, nthupzf}@gmail.com, mingyul@nvidia.com,sunmin@ee.nthu.edu.tw

Abstract

Most of the existing works on image description focus on generating expressive descriptions. The only few works that are dedicated to generating *stylish* (e.g., romantic, lyric, etc.) descriptions suffer from limited style variation and content digression. To address these limitations, we propose a controllable stylish image description generation model. It can learn to generate stylish image descriptions that are more related to image content and can be trained with the arbitrary monolingual corpus without collecting new paired image and stylish descriptions. Moreover, it enables users to generate various stylish descriptions by plugging in style-specific parameters to include new styles into the existing model. We achieve this capability via a novel layer normalization layer design, which we will refer to as the Domain Layer Norm (DLN). Extensive experimental validation and user study on various stylish image description generation tasks are conducted to show the competitive advantages of the proposed model.

Introduction

The image description generation (IDG) problem concerns about generating a natural language description that transcribes an input image. Over the years, tremendous effort has been dedicated to developing models that are descriptive. However, little effort is dedicated to generating descriptions that are *stylish* (e.g. romantic, lyric, etc). Even for the handful of stylish IDG models that exist, they only have a loose control over the style. Ideally, a stylish IDG model should allow users to flexibly control over the generated descriptions as shown in Fig 1. Such a model would be useful for increasing user engagement in applications requiring human interaction such as chatbot and social media sharing.

A naive approach to tackle the stylish IDG problem is to collect new corpora of paired images and descriptions for training. However, this is expensive. For each style that we wish to generate, we have to ask human annotators to write the romantic descriptions for each image in the training dataset.

In this paper, we propose a controllable stylish IDG model. Our model is jointly trained with a paired unstylish image description corpus (source domain) and a monolingual corpus of the specific style (target domain). In this setting, our model can learn to generate various styles without collecting new paired data in the target domain. Our main contribution is to show that the layer normalization can be used to disentangle language styles from the content of source and target domains via a small tweak. This design enables us to use the shared content to generate descriptions that are more relevant to the image as well as control the style by plugging in a set of style-specific parameters. We refer this mechanism as Domain Layer Normalization (DLN) since we treat each style as the target domain in the domain transfer setting.

We conduct an extensive experimental evaluation to validate the proposed approach using both subjective and objective performance metrics. We evaluate our model on four different styles, including fairy tale, romance, humor, and country song lyrics style (lyrics). Experiment results show that our model generates stylish descriptions that are more preferred by human subjects. It also outperforms prior works on the objective performance metrics.

GT (unstylish): A cat laying in a luggage bag on a bed
Romance: A cat laying in a luggage bag on a bed with his beloved owner every day. She just wake up, waiting for the kiss from owner.
Humorous: A cat laying in a luggage bag on a bed, thinking about to spend his life in that bag. But his owner is going to kick him out.
Lyrics: Funny cat scenes, on the front of my screen. Hey, my video on my table you know, you must know.
Fairy tale: A cat view himself as the king of the room, waiting for the servant to serve the food and water. He said: Food!

Figure 1: An ideal IDG can generate stylish descriptions for the given image. The generated descriptions should relate to the image content with different language styles.

Related Works

Visual style transfer. Image style transfer has been widely studied in computer vision. Gatys, Ecker, and Bethge (2015)

[*]equal contribution

synthesize a new stylish image by recombining image content with style features extracted from different images. Dumoulin, Shlens, and Kudlur (2017) propose to learn the style embedding of visual artistic style by conditioning on the parameter of batch normalization (Ioffe and Szegedy 2015). Huang and Belongie (2017) use adaptive instance norm. More recent approaches use the generative adversarial network (GAN) (Goodfellow et al. 2014) to align and transfer images from different domains. Liu and Tuzel (2016) employ weight-sharing assumption to learn the shared latent code between two domains and further propose translation stream in Liu, Breuel, and Kautz (2017) to encourage the same image in two domains to be mapped into common latent code. While our method is similar to these works in high level, the discrete property of text required new design.

Language style transfer. Supervised learning can be used to generate various linguistic attribute (e.g., different sentiments and different degrees of descriptiveness), but it requires a significant amount of labeled data. Many recent works assume there exist a share content space and a latent style vector between two non-parallel corpora for unsupervised language style transfer. Shen et al. (2017) propose an encoder-decoder structure with adversarial training to learning this space. Following the same line, Melnyk et al. (2017) introduce content preservation loss and classification loss to improve the transfer performance. Fu et al. (2018) propose to use a multi-decoder for different styles and a discriminator to learn a shared content code. Zhang (2018) also use similar structure by using shared and private encoder-decoder. In a recent work, Prabhumoye et al. (2018) introduce to ground the sentence in translation model, then apply adversarial training to get the desired style. What differs us from prior works is that we require generated stylish descriptions to match the visual content. Moreover, the style transferred in our work is more abstract instead of explicit styles such as sentiment, gender, or authorship in previous works.

Image description generation. Several works have been proposed to generate image descriptions by using paired image description data (Vinyals et al. 2015; Krause et al. 2017; Liang et al. 2017). To increase the naturalness and diversity of generated descriptions, Dai et al. (2017) apply adversarial training approach to train an evaluator to score the quality of generated descriptions. Chen et al. (2017) propose an adversarial training procedure to adapt image captioning style using unpaired images and captions. A new objective is proposed in Dai and Lin (2017) to enhance the distinctiveness of generated captions. On the other hand, there exist a few works proposed to enhance the attractiveness and style of the generated descriptions. Zhu et al. (2015) align the book and the corresponding movie release to a story-like description of the visual content. However, this method does not preserve the visual content. Mathews, Xie, and He (2016) propose the switch RNN to generate caption with positive and negative sentiments, which requires word level supervision and might not be able to scale. Recently, Gan et al. (2017b) investigate to generate tag-dependent caption by extending the weight matrix of LSTM to consider tag information. The following work StyleNet (Gan et al. 2017a) explores to decomposes LSTM matrix to incorporate the style information. One key difference is that we leverage an arbitrary stylish monolingual corpus that is not paired with any image dataset as target corpus instead of using paired images with stylish ground truth. Munigala et al. (2018) try to generate persuasive description for fashion image in unsupervise way but the descriptions are still transferred within fashion domain. The most similar to our work is Mathews, Xie, and He (2018), the major differences are that we do not exploit the language features such as POS tag of corpus and we do not pre-process the target corpus to make it similar to the source one. Our approach is end to end with minimal pre-process of target corpus.

Unsupervised Stylish Image Description Generation

The goal of stylish Image Description Generation (IDG) is to generate a natural language description d_T in space $\mathcal{D}_T$ given an image I in the image space $\mathcal{I}$. The style of the description is implicitly captured in the description space $\mathcal{D}_T$, where we use subscript T to emphasize the target style. There exist two settings for learning a stylish IDG model.

Supervised stylish IDG. In supervised stylish IDG, we are given a training dataset $\mathbb{D} = \{(I^{(n)}, d_T^{(n)}), n = 1, ..., N\}$, where each sample $(I^{(n)}, d_T^{(n)})$ is a pair of image and its target stylish description sampled from the joint distribution $p(\mathcal{I}, \mathcal{D}_T)$. The goal is to learn the conditional distribution $p(\mathcal{D}_T|\mathcal{I})$ using $\mathbb{D}$ so that we can generate stylish image descriptions for an input image.

Unsupervised stylish IDG. In unsupervised stylish IDG, we are given two training datasets $\mathbb{D}_S$ and $\mathbb{D}_T$. $\mathbb{D}_S = \{(I^{(n)}, d_S^{(n)}), n = 1, ..., N_S\}$ consists of pairs of image and its description $(I^{(n)}, d_S^{(n)})$ sampled from $p(\mathcal{I}, \mathcal{D}_S)$, where S is referred to as the source domain which is typically unstylish. $\mathbb{D}_T = \{(d_T^{(n)}), n = 1, ..., N_T\}$ is a dataset of target stylish descriptions $d_T^{(n)}$ sampled from $p(\mathcal{D}_T)$, where the corresponding images are not available. Hence, the learning task is considered as unsupervised. The goal of unsupervised stylish IDG is to learn the conditional distribution $p(\mathcal{D}_T|\mathcal{I})$ using $\mathbb{D}_S$ and $\mathbb{D}_T$.

Unsupervised stylish IDG is an ill-posed problem since it is about learning the conditional distribution $p(\mathcal{D}_T|\mathcal{I})$ without using samples from the joint distribution $p(\mathcal{I}, \mathcal{D}_T)$. Therefore, learning an unsupervised stylish IDG function is difficult without leveraging some useful assumptions. However, under the unsupervised setting, training data collection is greatly simplified: one could pair a general image description dataset (e.g., the MS-COCO dataset (Lin et al. 2014)) with an existing corpus of the target style (e.g., some romantic novels) for learning. A solution to the unsupervised problem could enable many stylish image description generation applications.

Unsupervised Stylish IDG via Domain Layer Norm

Assumptions. To deal with the ill-posed unsupervised stylish IDG problem, we make several assumptions illus-

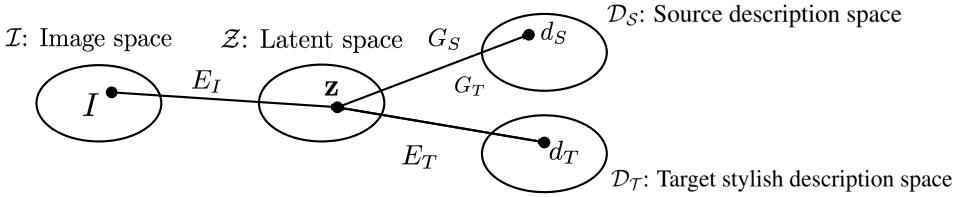

Figure 2: We make several assumptions to deal with the challenging unsupervised stylish image description generation problem. We first assume there exists a shared latent space $\mathcal{Z}$ so that a latent code $\mathbf{z} \in \mathcal{Z}$ can be mapped to the source description space $\mathcal{D}_S$ and the target stylish description space $\mathcal{D}_T$ via G_S and G_T. We also assume there exists a stylish image description embedding function E_T that can map a stylish description to a latent code. Finally, we assume there exists an image embedding function E_I that can map an image to a latent code. Once these functions are learned from data, we can generate a stylish image description for an image by applying E_I and G_T sequentially.

trated in Figure 2. We first assume that there exists a latent space $\mathcal{Z}$ providing a common ground to effectively map to and from the image space $\mathcal{I}$, the source description space $\mathcal{D}_S$, and the target stylish description space $\mathcal{D}_T$. From latent space to description space, we assume that there exists a source description generation function $G_S(\mathbf{z}) \in \mathcal{D}_S$ and a target stylish description generation function $G_T(\mathbf{z}) \in \mathcal{D}_T$. From non-latent space to latent pace, we assume that there exist an image encoder $E_I(I) \in \mathcal{Z}$ and a target description encoder $E_T(d_T) \in \mathcal{Z}$. Our goal is to learn the generation functions (G_T and G_S) and the encoding functions (E_I and E_T) from the unsupervised stylish IDG training data $\mathbb{D}_S$ and $\mathbb{D}_T$. Note that this is a challenging learning task if G_T and G_S is completely independent of each other. Hence, we assume that G_T and G_S share the ability to describe the same factual content but with different styles. Once these functions are learned, we can simply first encode the image I to a latent code using E_I and then using G_T to generate a stylish image description. In other words, the stylish image description is given by $G_T(E_I(I))$. We model the conditional distribution as $p(\mathcal{D}_T|\mathcal{I}) = \delta(G_T(E_I(I)))$, where δ is the delta function. Inspired by the success of deep learning, we model both of the generation and encoding functions using deep networks. Specifically, we model E_I using a deep convolutional neural network (CNN) (Krizhevsky, Sutskever, and Hinton 2012) and model E_T, G_T, and G_S using recurrent neural network as illustrated in Figure 3. We also use Skip-Thought Vectors (STV) (Kiros et al. 2015) to model E_T. For G_T and G_S, we use Layer Normalized Long Short Term Memory unit (LN-LSTM) as their recurrent module (Ba, Kiros, and Hinton 2016; Hochreiter and Schmidhuber 1997).

Training sketch. With the source domain dataset $\mathbb{D}_S$, we can train $\boldsymbol{z}_S = E_I(I)$ and $d_S = G_S(\boldsymbol{z}_S)$ jointly by solving the supervised IDG learning task, where $\boldsymbol{z}_S$ is the learned latent representation in the source domain. On the other hand, with the target domain dataset $\mathbb{D}_T$, we can train $\boldsymbol{z}_T = E_T(d_T)$ and $d_T = G_T(\boldsymbol{z}_T)$ jointly by solving an unsupervised description reconstruction learning task, where $\boldsymbol{z}_T$ is the learned latent representation in the target domain. To ensure that the latent space is shared (i.e., $\boldsymbol{z}_T \in \mathcal{Z}$ and $\boldsymbol{z}_S \in \mathcal{Z}$), we further assume that the generation functions G_S and G_T share most of their parameters.

Domain Layer Norm. Specifically, we assume G_S and G_T share all the parameters except those in their layer norm parameters (Ba, Kiros, and Hinton 2016). In other words, the domain description generators (G_S and G_T) only defer in the layer norm parameters. We refer this weight-sharing scheme as the Domain Layer Norm (DLN) scheme. The intuition behind DLN is to encourage the shared weight to capture the factual content between two domains while the differences (i.e., styles) are captured in layer norm parameters. This design helps G_T generate descriptions that are related to the image content even without the supervision of the corresponding images in training.

Training E_I and G_S via Supervised IDG. The goal of supervised image description generation is to learn $p(\mathcal{D}_S|\mathcal{I})$ by using $\mathbb{D}_S$. The G_S consists of an embedding matrix $\boldsymbol{\theta_W}$ that maps input text x_k to a vector $\boldsymbol{e}_k$, an LN-LSTM module, and an output matrix $\boldsymbol{\theta_V}$ that maps hidden state to predicted token $\hat{\boldsymbol{y}}$. Formally,

$$(\hat{\boldsymbol{y}}_{k+1}, \boldsymbol{h}_{k+1}) = G_S(\boldsymbol{e}_k, \boldsymbol{h}_k) \,, \tag{1}$$

$$\hat{\boldsymbol{y}}_{k+1} = \boldsymbol{\theta_V}^T \boldsymbol{h}_k \,, \tag{2}$$

$$\boldsymbol{e}_k = \boldsymbol{\theta_W}^T \mathbf{1}\{x_k\} \,, \tag{3}$$

$$\boldsymbol{e}_{-1} = E_I(I), \boldsymbol{h}_{-1} = \mathbf{0} \,, \tag{4}$$

where $\boldsymbol{h}_k$ is the hidden feature in the LN-LSTM, $k \in \{-1 \dots m-1\}$ is time step of description with length m, and $\mathbf{1}\{\}$ denotes the operator for one-hot encoding. To train the network, we minimize the sum of cross-entropy of correct words as follows,

$$\mathcal{L}_S = -\sum_{k=1}^{m} \log(\mathbf{1}\{x_k\}^T \hat{\boldsymbol{y}}_k) \,, \tag{5}$$

where x_k is the k^{th} word in the ground truth sentence.

Training E_T and G_T via Stylish Image Description Reconstruction. The G_T contains the LN-LSTM module, the same output matrix and embedding matrix used in G_S. For-

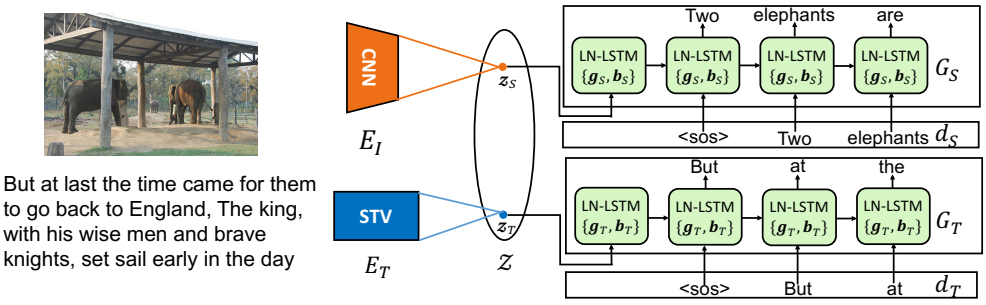

Figure 3: The E_I and E_T map the image and the target stylish description to a shared latent space. Both G_S and G_T share all weights except the layer norm parameters to capture the similar content in two domains. To disentangled the style factor, we employ different sets of layer norm parameters denoted as $\{\boldsymbol{g}_S, \boldsymbol{b}_S\}$ and $\{\boldsymbol{g}_T, \boldsymbol{b}_T\}$ for source and target domain during training.

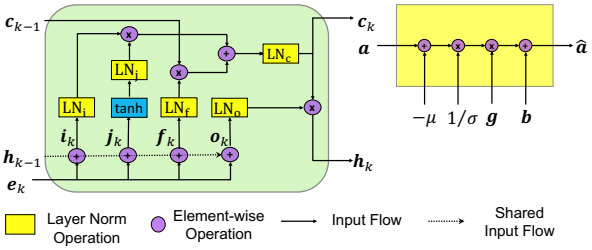

Figure 4: Inside the LN-LSTM cell (left) and the operation of layer normalization (right).

mally,

$$(\hat{\boldsymbol{y}}_{k+1}, \boldsymbol{h}_{k+1}) = G_T(\boldsymbol{e}_k, \boldsymbol{h}_k) \,, \tag{6}$$

$$\hat{\boldsymbol{y}}_{k+1} = \boldsymbol{\theta_V}^T \boldsymbol{h}_k \,, \tag{7}$$

$$\boldsymbol{e}_k = \boldsymbol{\theta_W}^T \mathbf{1}\{d_T^k\} \,, \tag{8}$$

$$\boldsymbol{e}_{-1} = E_T(d_T) \,, \tag{9}$$

$$\boldsymbol{h}_{-1} = \mathbf{0} \,, \tag{10}$$

where d_T is the target style image description. To train the network, we minimize the reconstruction error as follows,

$$\mathcal{L}_T = -\sum_{k=1}^{m} \log(\mathbf{1}\{d_T^k\}^T \hat{\boldsymbol{y}}_k) \,, \tag{11}$$

where d_T^k is the k^{th} word in the target style image description.

Relating G_S and G_T via Domain Layer Norm. We relate G_S and G_T by sharing all weights except layer norm parameters in the LN-LSTM. Details inside the LN-LSTM are shown in Fig 4, where the layer norm operation (LN) is applied to each gate of LSTM. Take the input gate as an example:

$$\hat{\boldsymbol{i}}_k = \mathrm{LN}(\boldsymbol{i}_k), \boldsymbol{i}_k = \boldsymbol{\theta}_{ie}\boldsymbol{e}_k + \boldsymbol{\theta}_{ih}\boldsymbol{h}_{k-1} \,, \tag{12}$$

where $\hat{\boldsymbol{i}}_k$ and $\boldsymbol{i}_k$ are the normalized and unnormalized input gates, $\boldsymbol{\theta}_{ie}$, $\boldsymbol{\theta}_{ih}$ are two projection matrices that map the embedding vector and the previous hidden state into the same dimension. The LN operation converts any input $\boldsymbol{a}$ to a normalized output $\hat{\boldsymbol{a}}$ as follows,

$$\hat{\boldsymbol{a}} = \frac{\boldsymbol{g}}{\sigma} \odot (\boldsymbol{a} - \mu)) + \boldsymbol{b} \,, \tag{13}$$

$$\mu = \frac{1}{p_h} \sum_{i=1}^{i=p_h} a_i \,, \tag{14}$$

$$\sigma = \sqrt{\frac{1}{p_h} \sum_{i=1}^{p_h} (a_i - \mu)^2} \,, \tag{15}$$

where a_i denotes the i^{th} entry in the vector $\boldsymbol{a}$, p_h is the dimention of the input $\boldsymbol{a}$, μ and σ are the mean and standard deviation of the input $\boldsymbol{a}$, $\boldsymbol{g}$ and $\boldsymbol{b}$ are scaling and shifting vectors (i.e., layer norm parameters) learned from the data.

We train the whole network by jointly minimizing the supervised IDG loss $\mathcal{L}_S$ and the unsupervised image description reconstruction loss $\mathcal{L}_T$ subject to the architectural constraint set to G_S and G_T as below, where λ is a hyperparameter.

$$\mathcal{L}(\boldsymbol{\theta}_{E_I}\boldsymbol{\theta}_{G_S}, \boldsymbol{\theta}_{E_T}, \boldsymbol{\theta}_{G_T}) = \lambda \mathcal{L}_S(\boldsymbol{\theta}_{E_I}, \boldsymbol{\theta}_{G_S}) + (1-\lambda)\mathcal{L}_T(\boldsymbol{\theta}_{E_T}, \boldsymbol{\theta}_{G_T}) \,. \tag{16}$$

Extension to New Target Styles. Given a model with parameters $\boldsymbol{\theta_V}$, $\boldsymbol{\theta_W}$, $\boldsymbol{\theta}_{E_I}$, and $\boldsymbol{\theta}_{G_S}$, pre-trained on a pair of the source and one target domain, we aim to adapt it to a new target domain (i.e, style) by enlarging $\boldsymbol{\theta_V}$ and $\boldsymbol{\theta_W}$ to $\boldsymbol{\theta}'_{\boldsymbol{V}}$ and $\boldsymbol{\theta}'_{\boldsymbol{W}}$ to accommodate new vocabulary and finetuning the remaining parameters to $\boldsymbol{\theta}'_{E_I}$, $\boldsymbol{\theta}'_{E_T}$, $\boldsymbol{\theta}'_{G_S}$ and $\boldsymbol{\theta}'_{G_T}$. Hence, we define a new loss function as:

$$\mathcal{L}(\boldsymbol{\theta}'_{E_I}, \boldsymbol{\theta}'_{G_S}, \boldsymbol{\theta}'_{E_T}, \boldsymbol{\theta}'_{G_T}) = \lambda_1 \mathcal{L}_S(\boldsymbol{\theta}'_{E_I}, \boldsymbol{\theta}'_{G_S}) + (1-\lambda_1)\mathcal{L}_T(\boldsymbol{\theta}'_{E_T}, \boldsymbol{\theta}'_{G_T}) + \lambda_2 R(\boldsymbol{\theta}'_{E_I}, \boldsymbol{\theta}'_{\boldsymbol{W}}, \boldsymbol{\theta}'_{\boldsymbol{V}}) \,, \tag{17}$$

where λ_1 and λ_2 are hyperparameters. The regularization term $R(\boldsymbol{\theta}'_{E_I}, \boldsymbol{\theta}'_{\boldsymbol{W}}, \boldsymbol{\theta}'_{\boldsymbol{V}}) = \|\boldsymbol{\theta}'_{E_I} - \boldsymbol{\theta}_{E_I}\|_2 + \|\boldsymbol{\theta}'_{\boldsymbol{W}} - \boldsymbol{\theta}_{\boldsymbol{W}}\|_2 + \|\boldsymbol{\theta}'_{\boldsymbol{V}} - \boldsymbol{\theta}_{\boldsymbol{V}}\|_2$ is used to prevent new weights from deviating the pretrained model. This encourages the adapted model to keep the information learned during the pretrained phase. We use pretrained $\boldsymbol{\theta}_{E_I}$ and $\boldsymbol{\theta}_{G_S}$ as initialization of $\boldsymbol{\theta}'_{E_I}$ and $\boldsymbol{\theta}'_{G_S}$. For $\boldsymbol{\theta}'_{G_T}$, we share all parameters in $\boldsymbol{\theta}'_{G_S}$ except the layer norm parameters. $\boldsymbol{\theta}'_{E_T}$ is trained from scratch. Note that we

do not update the source domain layer norm parameters since we do not need to learn source style.

Experiment

We conduct two experiments to evaluate our proposed method. First, we demonstrate that our method can generate stylish descriptions based on paired image and unstylish description in the source domain and a stylish monolingual corpus that is not paired with any image dataset in the target domain. Then, we demonstrate the flexibility of our DLN to progressively include new styles one by one in the second experiment. The implementation details are in the supplementary.

Evaluation Setting

Datasets. We use paragraphs released in (Krause et al. 2017) (VG-Para) as our source domain dataset. We do not use caption dataset such as MS-COCO because we found captions are less stylish when transfer to target style domain. We use pre-split data which contain 14575, 2489 and 2487 for training, validation and testing. For target dataset, we use humor and romance novel collections in BookCorpus (Zhu et al. 2015). We also collect country song lyrics and fairy tale to show that our method is effective on corpora with different syntactic structures and word usage. More details can be found in supplementary materials.

Baselines. We compare our method with four baselines: StyleNet (Gan et al. 2017a), Neural Story Teller (NST) (Kiros et al. 2015), DLN-RNN and Random. Stylenet generates stylish descriptions in an end-to-end way but with paired image and stylish ground truth description. NST breaks down the task into two steps, which first generate unstylish captions then apply style shift techniques to generate stylish descriptions. DLN-RNN uses the same framework as DLN with only difference in using simple recurrent neural network. Random samples the the same number of nouns as that in the unstylished ground truth from the corresponding vocabulary of target domain. Although a concurrent work (Mathews, Xie, and He 2018) that attempts to solve similar task as ours, the major differences are we do not exploit linguistic features and pre-process the target corpus to facilitate the training. Moreover, it is not sure whether the concurrent work can be applied to other styles or even multiple styles as it only makes a step toward generating sentences with romantic style.

Metrics of semantic relevance. As there is no ground truth sentences for stylish image descriptions in unpaired setting, the conventional n-gram based metrics such as BLEU (Papineni et al. 2002), METEOR (Denkowski and Lavie 2014) and CIDEr (Vedantam, Zitnick, and Parikh 2015) cannot be applied. It is also not suitable to calculate these metrics between stylish sentences and the unstylished ground truth because the goal of stylish description generation is to change the word usage while preserve certain semantic relevance between the stylish description and images.

We propose content similarity to evaluate the semantic relevance between generated stylish sentences and the unstylished ground truth. To calculate content similarity, we define C_S as the set of nouns in the ground truth (source domain), and C'_S as the union between C_S and synonyms for each noun in C_S, for the model may describe the same object with different words (e.g., cup and mug). Similar logic is applied to C_T and C'_T in the generated description (target domain). We calculate:

$$p = \frac{|C_T \cap C'_S|}{|C_T|} \qquad r = \frac{|C_S \cap C'_T|}{|C_S|}, \tag{18}$$

We take the f-score of the p and r as the content similarity score. The overall content similarity score is averaged over the testing data. This is because we assume stylish descriptions should at least contain objects which appear in the image. We also report SPICE (Anderson et al. 2016) score, which calculate the f-score of semantic tuples between untylished ground truth and the generated stylish descriptions. The final score is average over all testing data.

Metrics of stylishness. We use transfer accuracy to evaluate the stylishness of our generated description. The transfer accuracy is widely used in language style transfer task (Shen et al. 2017; Melnyk et al. 2017; Fu et al. 2018). It measures how often do descriptions have labels of target style on test dataset based on a pre-trained style classifier. We follow the definition of transfer accuracy in (Fu et al. 2018), which is

$$\mathcal{T} = \begin{cases} 1 & \text{if } s > 0.5 \\ 0 & \text{if } s \leq 0.5 \end{cases} \tag{19}$$

where s is the output probability score of the classifier. We define $R_T = \frac{N_{vt}}{N_{vs}}$ as our transfer accuracy, which is the fraction of number of testing N_{vs} data in source domain and number of testing data that correctly transfer description with target style N_{vt}. The final score is average over all testing data.

Human evaluation. The difficulty in generating stylish sentence in unpaired setting is to remain semantic relevance. Therefore, we conduct a human study on Amazon Mechanical Turk (AMT) independently for each methods to judge the semantic relevance between image and description. For each model, we randomly sample 100 images then generate stylish descriptions for each style. Two workers are asked to vote the semantic relevance with following prompt: Given an image and a paragraph from the book (Our stylish corpus), how well does the paragraph content relate to objects in the image. Workers are forced to vote from unrelated to related. The criteria for eligible workers are having at least 100 successful HITs with 70% acceptance rate. The total number of HIT is 2400. For each HIT, the order of options is randomized. Workers are forced to vote and all responses are counted without aggregation.

Results

The result of the first experiment is summarized in Table 1. We also report p, r and the numerator of each for further comparison. It is worth noting that the perfect transfer accuracy may not be the best since the model could greedily generate the vocabulary used in the target domain and digress from the image content. Therefore, an ideal stylish description is the one with the high content similarity score and an acceptable transfer accuracy. Our DLN consistently outperforms other baselines in term of all semantic related metrics with a

Model	Data	CS	S	T	p	r	n_p	n_r
NST (Kiros et al. 2015)	Lyrics	0.037	0.016	100%	0.041	0.044	0.68	0.75
StyleNet (Gan et al. 2017a)	Lyrics	0.033	0.014	100%	0.038	0.038	0.57	0.67
Random	Lyrics	0.008	0.002	55.2%	0.007	0.012	0.13	0.09
DLN-RNN	Lyrics	0.072	0.030	100%	**0.101**	0.069	**1.65**	1.17
DLN	Lyrics	**0.083**	**0.033**	99.2%	0.080	**0.115**	1.25	**1.92**
NST (Kiros et al. 2015)	Romance	0.088	0.039	100%	0.087	0.113	**1.57**	1.90
StyleNet (Gan et al. 2017a)	Romance	0.012	0.005	100%	0.032	0.001	0.11	0.14
Random	Romance	0.005	0.002	100%	0.004	0.001	0.07	0.05
DLN-RNN	Romance	0.083	0.034	94.3%	0.078	0.125	1.27	0.71
DLN	Romance	**0.151**	**0.058**	95.4%	**0.193**	**0.148**	1.56	**2.43**
NST (Kiros et al. 2015)	Humor	0.103	0.041	99.7%	0.097	0.143	2.22	2.44
StyleNet (Gan et al. 2017a)	Humor	0.010	0.005	99.8%	0.024	0.001	0.12	0.15
Random	Humor	0.007	0.002	100%	0.006	0.014	0.11	0.07
DLN-RNN	Humor	0.093	0.038	89.5%	0.095	0.12	1.58	0.92
DLN	Humor	**0.173**	**0.065**	70.0%	**0.205**	**0.182**	**2.32**	**2.99**
NST (Kiros et al. 2015)	Fairy tale	0.116	0.044	99.8%	0.116	**0.145**	**2.47**	**2.44**
StyleNet (Gan et al. 2017a)	Fairy tale	0.028	0.013	99.8%	0.045	0.026	0.34	0.46
Random	Fairy tale	0.004	0.001	100%	0.003	0.010	0.06	0.04
DLN-RNN	Fairy tale	0.084	0.033	79.5%	0.076	0.140	1.22	0.72
DLN	Fairy tale	**0.135**	**0.050**	93.7%	**0.194**	0.125	1.29	2.06

Table 1: Performance comparison between DLN and several baselines. CS, S and T stand for content similarity, SPICE and transfer accuracy. p and r are as defined in Eq. 18. n_p and n_r are the numerator of each. DLN has generally higher score of content related metrics. Higher is better for all metrics except the transfer accuracy.

marginal drop of transfer accuracy on most datasets. All baselines are better than Random, which suggests all baselines can generate semantic-related description to certain degree. We observe NST has large n_p and n_r in fairy tale. We think this is because NST tends to generate long sentences. For each style (Fairy, Humor, Romance, and Lyrics), the average sentence length of NST is $(119, 109, 103, 84)$ while that of DLN is $(38, 54, 41, 97)$. Therefore, it is possible that NST generates more nouns in the unstylish ground truth.

We also report the performance of DLN and DLN-RNN on unstylish description generation task in Table 2. We calculate the BLEU-4, METEOR and CIDEr scores between generated sentences and unstylished ground truth. Combined with the result of stylish description generation in Table 1, we can conclude that the proposed domain layer norm can benefit the unpaired image to stylish description as we have a better model in conventional image to text generation.

The result of human study is shown in Fig 5, we report the best of our model in Table 1 (DLN) and other baselines for comparison. The DLN has the highest related and lowest unrelated votes while over half of descriptions are voted as unrelated in other baselines. Qualitative results in Fig 6 shows that the description generated by DLN is related to images. Note that the goal of generated stylish description is not to match every factual aspect of images, it should better be judged whether the description is related to the image if the image appears in the target corpus.

Multi-style. We progressively expand DLN to include three target domains (fairy, romance, lyrics) to demonstrate the flexibility of our model. In other words, we follow Eq 17 to train source and fairy tale style then include romance and lyrics style, which is denoted as DLN-Multi. To generate the description, we use the same target decoder with a different style-specific embedding matrix, layer norm parameters, and output matrix. We conduct another human study by asking five workers to determine the best description given following

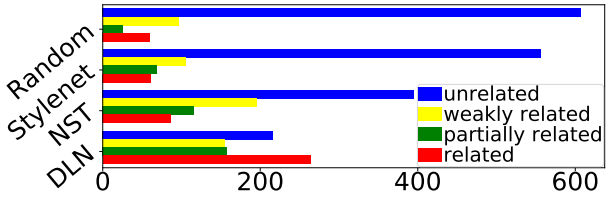

Figure 5: Human study of semantic relevance of all methods. DLN has highest related and lowest unrelated votes compared to other baselines.

Model	BLEU-3	BLEU-4	METEOR	CIDEr
DLN-RNN	0.106	0.062	0.130	0.069
DLN	**0.132**	**0.080**	**0.150**	**0.127**

Table 2: Performance on generate unstylish description. DLN is better than DLN-RNN in all metrics.

Fairy Tale

GT (unstylish): An Asian woman laughs while looking at the martini glass of an Asian man. Both people have black hair. The man is dressed in a collared shirt and a black tie. He is holding his glass up in the air. The woman is wearing a black sleeveless dress.
DLN: A man and woman were having a large party at their own time. But the woman made them very uneasy, for he thought that they might take all their money and as a man of gold.

Lyrics

GT (unstylish): Grey and white birds are flying in the air. The sky is blue and have some white clouds. On the background there is brown hills covered with white snow. The bird on the foreground has the wings extended and flies to the right. The bird has a long black beck. The ground is wet and muddy and is covered with stones.
DLN: It's deep blue; Like a bird that is flying high; This is the heart and not a soul to find; My life is a sin made of stone; Just look what I see, what I want to know; Wherever are you; I want to be loved by you; I want to be loved by you; I want to be loved and you left me to be alone; A true and little I gave you all my heart; I wish you could change your mind;

Humor

GT (unstylish): A man stands on top of a snow-covered mountain. Behind him are a bunch of other mountains. These mountains are covered in snow as well. The man at the top of the mountain is strapped into a pair of skis and is wearing grey ski boots. He has on black snow pants, a blue and grey jacket, black sunglasses and a blue beanie.
DLN: A man had said that I was going to die on the mountain! Then now he was tied to his feet. No, no matter, no running. He was dressed in perfect sweat and shirt and jumping in the air.

Romance

GT (unstylish): There are three people in the living room. One person is sitting on the gray couch. The other two people are standing up in the middle of the room, in front of the person on the couch. The two people that are standing have a game controller in their hands. They are all facing in the same direction and have smiles on their faces.
DLN: People sitting in a room together, and Mom and Dad sitting across the hall from the kitchen floor, laughing, something good friends looking over me and let alone a room. You're in love with me tonight!

Figure 6: Examples of stylish descriptions by DLN. Note the goal of stylish description is not to match every factual aspect of the image. It should be better judged whether the descriptions are related to the image if the image appears in the context of the target corpus. The semicolon (;) in lyrics serves as new line symbol.

priorities: content, style, and naturalness. This prompt forces workers to choose the better one if the two options are equally related to images. We sample 100 images for each and use the same criteria to select workers. The result is presented in Table 3, which shows the performance of DLN-Multi is competitive to DLN. DLN-Multi thus gives users the capability to include new style into the existing model, which is a novel feature not reported in other baselines.

Discussion: transfer accuracy and domain shift. We observe a drop in transfer accuracy on the source to humor transfer in DLN, and we believe this is related to the scale of domain shift. To quantify this, we analyze the percentage of shared noun between the source ($V_{src} = 6.2$k) and target domain, which are $(50\%, 68\%, 74\%, 60\%)$ for lyrics, romance humor and fairy tale. For the transfer from the source to humor domain, the shared nouns account for over 70% nouns in the source domain, which means the domain shift between the source and humor is smaller than others. This makes it more difficult for the classifier to distinguish two domains. Therefore, the transfer accuracy of the source to humor is lower. We note Random get lowest transfer accuracy in lyrics style and we believe this is because sampling word from the vocabulary of lyrics alone cannot have sentences with new line symbol (i.e. ;), which is an important feature for being classified as stylish.

Conclusion and future work

We propose a novel unsupervised stylish IDG model via domain layer norm with the capability to progressively include new styles. Experiment results show that our stylish IDG results are more preferred by human subjects. We plan to invesitgate the intermediate style generated by interpolation of domain layer norm parameter and address the fluency of generated sentences in the future.

Model	Style	CS	S	T	P
DLN-Multi	Romance	0.116	0.047	97.1%	36.7%
DLN	Romance	**0.151**	0.058	95.4%	**63.3%**
DLN-Multi	Lyrics	**0.118**	0.047	99.7%	**54.3%**
DLN	Lyrics	0.083	0.033	99.2%	45.8%
DLN-Multi	Fairy tale	0.120	0.048	99.0%	47.4%
DLN	Fairy tale	**0.135**	0.050	93.7%	**52.6%**

Table 3: Result of DLN and DLN-Multi. CS, S, T and P are content similarity, SPICE, tansfer accuracy and human preference score. Overall, the performance of DLN-Multi is competitive to DLN in all metrics.

Acknowledgement We thanks Nvidia for collaboration. We also thanks MOST-107 2634-F-007-007, MOST Joint Research Center for AI Technology and All Vista Healthcare and MediaTeck for their support.

References

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, 382–398. Springer.

Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. In *Advances in Neural Information Processing Systems (NIPS)*.

Chen, T.-H.; Liao, Y.-H.; Chuang, C.-Y.; Hsu, W.-T.; Fu, J.; and Sun, M. 2017. Show, adapt and tell: Adversarial training of cross-domain image captioner. In *IEEE International Conference on Computer Vision (ICCV)*.

Dai, B., and Lin, D. 2017. Contrastive learning for image captioning. In *Advances in Neural Information Processing Systems (NIPS)*.

Dai, B.; Lin, D.; Urtasun, R.; and Fidler, S. 2017. Towards diverse and natural image descriptions via a conditional gan. In *IEEE International Conference on Computer Vision (ICCV)*.

Denkowski, M., and Lavie, A. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, 376–380.

Dumoulin, V.; Shlens, J.; and Kudlur, M. 2017. A learned representation for artistic style.

Fu, Z.; Tan, X.; Peng, N.; Zhao, D.; and Yan, R. 2018. Style transfer in text: Exploration and evaluation.

Gan, C.; Gan, Z.; He, X.; Gao, J.; and Deng, L. 2017a. Stylenet: Generating attractive visual captions with styles. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gan, Z.; Gan, C.; He, X.; Pu, Y.; Tran, K.; Gao, J.; Carin, L.; and Deng, L. 2017b. Semantic compositional networks for visual captioning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2015. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*.

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 (NIPS)*.

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

Huang, X., and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE International Conference on Computer Vision (ICCV)*.

Ioffe, S., and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*.

Kiros, R.; Zhu, Y.; Salakhutdinov, R. R.; Zemel, R.; Urtasun, R.; Torralba, A.; and Fidler, S. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems (NIPS)*.

Krause, J.; Johnson, J.; Krishna, R.; and Fei-Fei, L. 2017. A hierarchical approach for generating descriptive image paragraphs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*.

Liang, X.; Hu, Z.; Zhang, H.; Gan, C.; and Xing, E. P. 2017. Recurrent topic-transition gan for visual paragraph generation. In *IEEE International Conference on Computer Vision (ICCV)*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*.

Liu, M.-Y., and Tuzel, O. 2016. Coupled generative adversarial networks. In *Advances in Neural Information Processing Systems (NIPS)*.

Liu, M.-Y.; Breuel, T.; and Kautz, J. 2017. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems (NIPS)*.

Mathews, A. P.; Xie, L.; and He, X. 2016. Senticap: Generating image descriptions with sentiments.

Mathews, A.; Xie, L.; and He, X. 2018. Semstyle: Learning to generate stylised image captions using unaligned text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8591–8600.

Melnyk, I.; Santos, C. N. d.; Wadhawan, K.; Padhi, I.; and Kumar, A. 2017. Improved neural text attribute transfer with non-parallel data.

Munigala, V.; Mishra, A.; Tamilselvam, S. G.; Khare, S.; Dasgupta, R.; and Sankaran, A. 2018. Persuaide! an adaptive persuasive text generation system for fashion domain. In *Companion of the The Web Conference 2018 on The Web Conference 2018*, 335–342. International World Wide Web Conferences Steering Committee.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318. Association for Computational Linguistics.

Shen, T.; Lei, T.; Barzilay, R.; and Jaakkola, T. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems (NIPS)*.

Prabhumoye, S.; Tsvetkov, Y.; Salakhutdinov, R.; and Black, A. W. 2018. Style transfer through back-translation. In *Association for Computational Linguistics (ACL)*.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

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)*.

Ye Zhang, Nan Ding, R. S. 2018. Shaped: Shared-private encoder-decoder for text style adaptation. In *The North American Chapter of the Association for Computational Linguistics (NAACL)*.

Zhu, Y.; Kiros, R.; Zemel, R.; Salakhutdinov, R.; Urtasun, R.; Torralba, A.; and Fidler, S. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *IEEE International Conference on Computer Vision (ICCV)*.