

# Unsupervised Stylish Image Description Generation via Domain Layer Norm

**Cheng-Kuan Chen<sup>†\*</sup>, Zhu Feng Pan<sup>†\*</sup>, Min Sun<sup>†</sup>, Ming-Yu Liu<sup>‡</sup>**  
 National Tsing-Hua University<sup>†</sup>, Nvidia<sup>‡</sup>

## 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



**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.

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.

\*Equal contribution

Copyright © 2019, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

## Related Works

**Visual style transfer.** Image style transfer has been widely studied in computer vision. Gatys et al. (Gatys, Ecker, and Bethge 2015) synthesize a new stylish image by recombining image content with style features extracted from different images. Dumoulin et al. (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 et al. (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 et al. (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 language required new model 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. (Shen et al. 2017) propose an encoder-decoder structure with adversarial training to learning this space. Following the same line, Melnyk et al. (Melnyk et al. 2017) introduce content preservation loss and classification loss to improve the transfer performance. Fu et al. (Fu et al. 2018) propose to use a multi-decoder for different styles and a discriminator to learn a shared content code. Zhang et al. (Ye Zhang 2018) also use similar structure by using shared and private encoder-decoder. In a recent work, Prabhumoye et al. (Shrimai Prabhumoye 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. (Dai et al. 2017) apply adversarial training approach to train an evaluator to score the quality of generated descriptions. Chen et al. (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. (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 con-

tent. Matthews et al. (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. (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. 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  $\mathbb{D} = \{(I^{(n)}, d_T^{(n)}), n = 1, \dots, N\}$ , 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  $\mathbb{D}_T$ .  $\mathbb{D}_S = \{(I^{(n)}, d_S^{(n)}), n = 1, \dots, N_S\}$  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.  $\mathbb{D}_T = \{(d_T^{(n)}), n = 1, \dots, N_T\}$  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 illustrated 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  $\delta$  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  $\mathbf{z}_S = E_I(I)$  and  $d_S = G_S(\mathbf{z}_S)$  jointly by solving the supervised IDG learning task, where  $\mathbf{z}_S$  is the learned latent representation in the source domain. On the other hand, with the target domain dataset  $\mathbb{D}_T$ , we can train  $\mathbf{z}_T = E_T(d_T)$  and  $d_T = G_T(\mathbf{z}_T)$  jointly by solving an unsupervised description reconstruction learning task, where  $\mathbf{z}_T$  is the learned latent representation in the target domain. To ensure that the latent space is shared (i.e.,  $\mathbf{z}_T \in \mathcal{Z}$  and  $\mathbf{z}_S \in \mathcal{Z}$ ), 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 cap-

ture 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  $\theta_W$  that maps input text  $x_k$  to a vector  $e_k$ , an LN-LSTM module, and an output matrix  $\theta_V$  that maps hidden state to predicted token  $\hat{y}$ . Formally,

$$(\hat{y}_{k+1}, \mathbf{h}_{k+1}) = G_S(e_k, \mathbf{h}_k), \quad (1)$$

$$\hat{y}_{k+1} = \theta_V^T \mathbf{h}_k, \quad (2)$$

$$e_k = \theta_W^T \mathbf{1}\{x_k\}, \quad (3)$$

$$e_{-1} = E_I(I), \mathbf{h}_{-1} = \mathbf{0}, \quad (4)$$

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

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

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  $G_S$ . Formally,

$$(\hat{y}_{k+1}, \mathbf{h}_{k+1}) = G_T(e_k, \mathbf{h}_k), \quad (6)$$

$$\hat{y}_{k+1} = \theta_V^T \mathbf{h}_k, \quad (7)$$

$$e_k = \theta_W^T \mathbf{1}\{d_T^k\}, \quad (8)$$

$$e_{-1} = E_T(d_T), \quad (9)$$

$$\mathbf{h}_{-1} = \mathbf{0}, \quad (10)$$

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

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

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:

$$\hat{i}_k = \text{LN}(i_k), i_k = \theta_{ie} e_k + \theta_{ih} \mathbf{h}_{k-1}, \quad (12)$$

where  $\hat{i}_k$  and  $i_k$  are the normalized and unnormalized input gates,  $\theta_{ie}$ ,  $\theta_{ih}$  are two projection matrices that map the embedding vector and the previous hidden state into the same

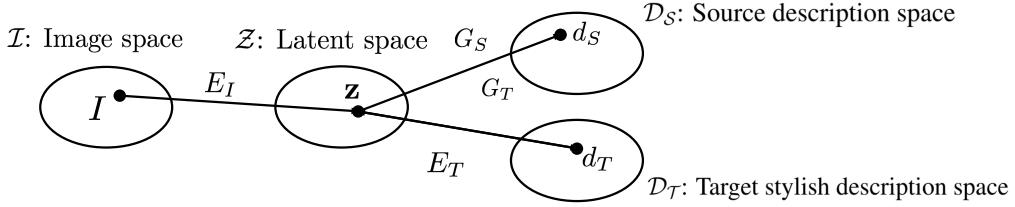


Figure 2: We make several assumptions to deal with the challenging unsupervised stylish image description generation problem. We first assume there exists a shared latent space  $\mathcal{Z}$  so that a latent code  $\mathbf{z} \in \mathcal{Z}$  can be mapped to the source description space  $\mathcal{D}_S$  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.

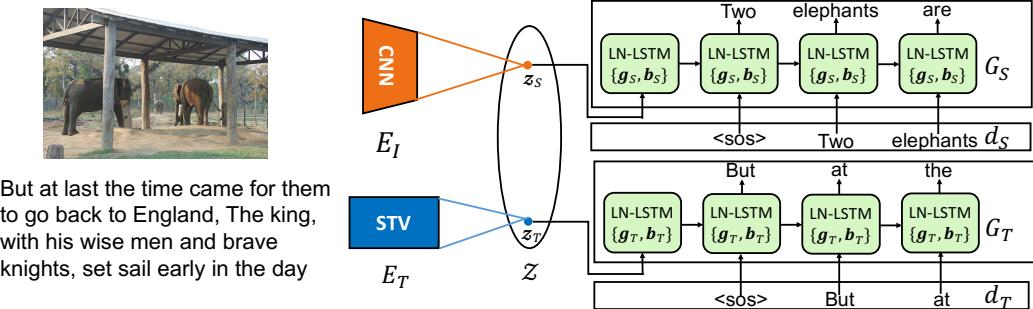


Figure 3: The  $E_I$  and  $E_T$  map the image and the target stylish description to a shared latent space. Both  $G_S$  and  $G_T$  share all weights except the layer norm parameters to capture the similar content in two domains. To disentangled the style factor, we employ different sets of layer norm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

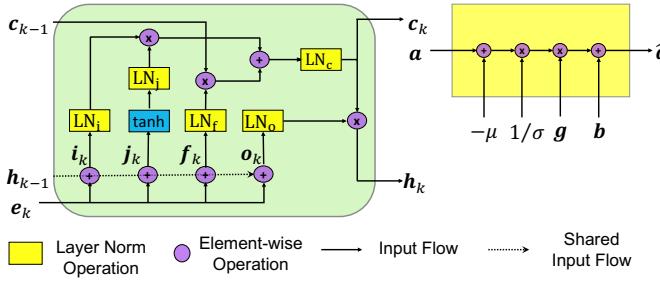


Figure 4: Inside the LN-LSTM cell (left) and the operation of layer normalization (right).

dimension. The LN operation converts any input  $\mathbf{a}$  to a normalized output  $\hat{\mathbf{a}}$  as follows,

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

$$\mu = \frac{1}{p_h} \sum_{i=1}^{i=p_h} a_i, \quad (14)$$

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

where  $a_i$  denotes the  $i^{th}$  entry in the vector  $\mathbf{a}$ ,  $p_h$  is the di-

mention of the input  $\mathbf{a}$ ,  $\mu$  and  $\sigma$  are the mean and standard deviation of the input  $\mathbf{a}$ ,  $\mathbf{g}$  and  $\mathbf{b}$  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  $\lambda$  is a hyperparameter.

$$\begin{aligned} \mathcal{L}(\theta_{E_I}, \theta_{G_S}, \theta_{E_T}, \theta_{G_T}) &= \lambda \mathcal{L}_S(\theta_{E_I}, \theta_{G_S}) \\ &\quad + (1 - \lambda) \mathcal{L}_T(\theta_{E_T}, \theta_{G_T}). \end{aligned} \quad (16)$$

**Extension to New Target Styles.** Given a model with parameters  $\theta_V$ ,  $\theta_W$ ,  $\theta_{E_I}$ , and  $\theta_{G_S}$ , 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  $\theta_V$  and  $\theta_W$  to  $\theta'_V$  and  $\theta'_W$  to accommodate new vocabulary and finetuning the remaining parameters to  $\theta'_{E_I}$ ,  $\theta'_{E_T}$ ,  $\theta'_{G_S}$  and  $\theta'_{G_T}$ . Hence, we define a new loss function as:

$$\begin{aligned} \mathcal{L}(\theta'_{E_I}, \theta'_{G_S}, \theta'_{E_T}, \theta'_{G_T}) &= \lambda_1 \mathcal{L}_S(\theta'_{E_I}, \theta'_{G_S}) \\ &\quad + (1 - \lambda_1) \mathcal{L}_T(\theta'_{E_T}, \theta'_{G_T}) + \lambda_2 R(\theta'_{E_I}, \theta'_W, \theta'_V), \end{aligned} \quad (17)$$

where  $\lambda_1$  and  $\lambda_2$  are hyperparameters. The regularization term  $R(\theta'_{E_I}, \theta'_W, \theta'_V) = \|\theta'_{E_I} - \theta_{E_I}\|_2 + \|\theta'_W - \theta_W\|_2 + \|\theta'_V - \theta_V\|_2$  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  $\theta_{E_I}$  and  $\theta_{G_S}$  as initialization of  $\theta'_{E_I}$  and  $\theta'_{G_S}$ . For  $\theta'_{G_T}$ , we share all parameters in  $\theta'_{G_S}$  except the layer norm parameters.  $\theta'_{E_T}$  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, Lawrence 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:

$$p = \frac{|C_T \cap C'_S|}{|C_T|} \quad r = \frac{|C_S \cap C'_T|}{|C_S|}, \quad (18)$$

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 unstylished ground truth and the generated stylish descriptions. The final score is average over all testing data.

**Metrics of styliness.** We use transfer accuracy to evaluate the styliness 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

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

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

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)	0.033	0.014	100%	0.038	0.038	0.57	0.67
	Random	0.008	0.002	55.2%	0.007	0.012	0.13	0.09
	DLN-RNN	0.072	0.030	100%	<b>0.101</b>	0.069	<b>1.65</b>	1.17
	DLN	<b>0.083</b>	<b>0.033</b>	99.2%	0.080	<b>0.115</b>	1.25	<b>1.92</b>
NST (Kiros et al. 2015)	Romance	0.088	0.039	100%	0.087	0.113	<b>1.57</b>	1.90
	StyleNet (Gan et al. 2017a)	0.012	0.005	100%	0.032	0.001	0.11	0.14
	Random	0.005	0.002	100%	0.004	0.001	0.07	0.05
	DLN-RNN	0.083	0.034	94.3%	0.078	0.125	1.27	0.71
	DLN	<b>0.151</b>	<b>0.058</b>	95.4%	<b>0.193</b>	<b>0.148</b>	1.56	<b>2.43</b>
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)	0.010	0.005	99.8%	0.024	0.001	0.12	0.15
	Random	0.007	0.002	100%	0.006	0.014	0.11	0.07
	DLN-RNN	0.093	0.038	89.5%	0.095	0.12	1.58	0.92
	DLN	<b>0.173</b>	<b>0.065</b>	70.0%	<b>0.205</b>	<b>0.182</b>	<b>2.32</b>	<b>2.99</b>
NST (Kiros et al. 2015)	Fairy tale	0.116	0.044	99.8%	0.116	<b>0.145</b>	<b>2.47</b>	<b>2.44</b>
	StyleNet (Gan et al. 2017a)	0.028	0.013	99.8%	0.045	0.026	0.34	0.46
	Random	0.004	0.001	100%	0.003	0.010	0.06	0.04
	DLN-RNN	0.084	0.033	79.5%	0.076	0.140	1.22	0.72
	DLN	<b>0.135</b>	<b>0.050</b>	93.7%	<b>0.194</b>	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 of transfer accuracy.

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 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 unstylized 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

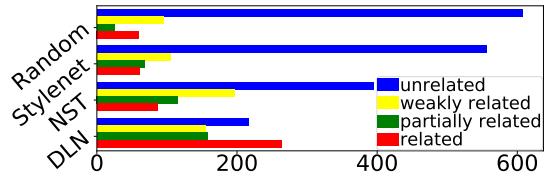


Figure 5: Human study of semantic relevance of all methods. DLN has highest related and lowest unrelated votes compared to other baselines.

Model	BLEU-3	BLEU-4	METEOR	CIDEr
DLN-RNN	0.106	0.062	0.130	0.069
DLN	<b>0.132</b>	<b>0.080</b>	<b>0.150</b>	<b>0.127</b>

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

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

## 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.

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 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 ( $V_{src} = 6.2k$ ) 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

Model	Style	CS	S	T	P
DLN-Multi	Romance	0.116	0.047	97.1%	36.7%
DLN	Romance	<b>0.151</b>	0.058	95.4%	<b>63.3%</b>
DLN-Multi	Lyrics	<b>0.118</b>	0.047	99.7%	<b>54.3%</b>
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	<b>0.135</b>	0.050	93.7%	<b>52.6%</b>

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.

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 investitgate the intermediate style generated by interpolation of domain layer norm parameter and address the fluency of generated sentences in the future.

## References

- [Abadi et al. 2016] Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. 2016. Tensorflow: A system for large-scale machine learning. In *USENIX Symposium on Operating Systems Design and Implementation (OSDI)*.
- [Anderson et al. 2016] 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, Kiros, and Hinton 2016] Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. In *Advances in Neural Information Processing Systems (NIPS)*.
- [Chen et al. 2017] 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 and Lin 2017] Dai, B., and Lin, D. 2017. Contrastive learning for image captioning. In *Advances in Neural Information Processing Systems (NIPS)*.
- [Dai et al. 2017] 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 and Lavie 2014] 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, Shlens, and Kudlur 2017] Dumoulin, V.; Shlens, J.; and Kudlur, M. 2017. A learned representation for artistic style.
- [Fu et al. 2018] Fu, Z.; Tan, X.; Peng, N.; Zhao, D.; and Yan, R. 2018. Style transfer in text: Exploration and evaluation.
- [Gan et al. 2017a] 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 et al. 2017b] 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, Ecker, and Bethge 2015] Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2015. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*.
- [Goodfellow et al. 2014] 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 and Schmidhuber 1997] Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.
- [Huang and Belongie 2017] 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 and Szegedy 2015] Ioffe, S., and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*.
- [Kim 2014] Kim, Y. 2014. Convolutional neural networks for sentence classification. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- [Kingma and Ba 2015] Kingma, D. P., and Ba, J. 2015. Adam: A method for stochastic optimization.
- [Kiros et al. 2015] 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)*.
- [Kiros, Salakhutdinov, and Zemel 2014] Kiros, R.; Salakhutdinov, R.; and Zemel, R. S. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*.
- [Krause et al. 2017] 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, Sutskever, and Hinton 2012] 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 et al. 2017] 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 et al. 2014] 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 and Tuzel 2016] Liu, M.-Y., and Tuzel, O. 2016. Coupled generative adversarial networks. In *Advances in Neural Information Processing Systems (NIPS)*.
- [Liu, Breuel, and Kautz 2017] Liu, M.-Y.; Breuel, T.; and Kautz, J. 2017. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems (NIPS)*.
- [Mathews, Xie, and He 2016] Mathews, A. P.; Xie, L.; and He, X. 2016. Senticap: Generating image descriptions with sentiments.
- [Mathews, Xie, and He 2018] 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 et al. 2017] Melnyk, I.; Santos, C. N. d.; Wadhawan, K.; Padhi, I.; and Kumar, A. 2017. Improved neural text attribute transfer with non-parallel data.
- [Papineni et al. 2002] 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 et al. 2017] 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)*.
- [Shrimai Prabhumoye 2018] Shrimai Prabhumoye, Yulia Tsvetkov, R. S. A. W. B. 2018. Style transfer through back-translation. In *Association for Computational Linguistics (ACL)*.
- [Vedantam, Lawrence Zitnick, and Parikh 2015] 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 et al. 2015] 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 2018] 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 et al. 2015] 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)*.

## Appendix

The content of this supplementary material is summarized as below:

- Data statistic
- Implementation details of baselines and DLN
- Interface of human evaluation
- More qualitative examples

### Data Statistic

For romance and humor data, we randomly sampled 50000 passages from BookCorpus (Zhu et al. 2015). For fairy tale, we crawled from the website<sup>1</sup> and also sample 50000 passages from it. For the country song lyrics, we use the data released by Kaggle<sup>2</sup> and use all country song lyrics as our corpus. The Fig 7 shows the word usage of each corpus.

### Baselines

- **Neural Story Teller** (Kiros et al. 2015): NST contains two separate modules: image to caption in the source domain and stylish story decoder in the target domain. The first module is used to extract the source domain textual representation of the image. They use pre-trained image caption alignment (Kiros, Salakhutdinov, and Zemel 2014) on MS-COCO to extract the top  $N$  neighbor captions for an image. The textual representation of the image is calculated by averaging the skip-thought encoded vector of top  $N$  captions. The second module is used to reconstruct target domain textual representation to original stylish story passage, where a decoder is trained to reconstruct the textual representation of stylish story passage to the original stylish story passage. The textual representation here is also the skip-thought encoded vector. To generate the description for given image, it first subtracts the mean of all skip-thought encoded MS-COCO caption and added the mean of skip-thought encoded stylish story passage to transform the source textual representation to the target one; then, feed it into the second module to generate the stylish story.
- **StyleNet** (Gan et al. 2017a): We re-implement StyleNet as one of our baseline. The StyleNet is based on factorized LSTM, which factorized the weights mapping inputs to hidden representation to  $\theta_W = USV$ , where  $U \in \mathbb{R}^{p_e \times p_m}$ ,  $S \in \mathbb{R}^{p_m \times p_n}$  and  $V \in \mathbb{R}^{p_n \times p_h}$ , where  $p_e, p_h$  are dimension of embedding and hidden size, and  $p_m, p_n$  are the dimension of the matrix. It contains a language model trained on target corpus and a source image to caption decoder where each factorized LSTM is associated with a style matrix  $S_s$  and  $S_t$ . During inference, it replace the source style matrix with target style matrix and generate descriptions.

### Implementation Details

**NST.** We follow the original setting and implementation<sup>3</sup> and train the decoder on our corpus till converge. We set the maximum length of text data equals to 100.

**DLN.** We use the same text length as used in NST. We use most common 10000 vocabularies in the source domain. For target domain, we use 10000 vocabularies for lyrics and 15000 vocabularies for romance, humor and fairy tale corpora.

<sup>1</sup><http://www.loyalbooks.com>

<sup>2</sup><https://www.kaggle.com/gyanis95/380000-lyrics-from-metrolyrics>

<sup>3</sup><https://github.com/ryankiros/neural-storyteller>

We use skips-thought vector released by Tensorflow (Abadi et al. 2016) as our  $E_T$ . We follow the original NST implementation to stack uni-skip and bi-skip skips-thought vector to get 4800 dimension feature for text. We use pre-split training set in VG-Para (Krause et al. 2017).

We use ResNet50 from Keras<sup>4</sup> as our  $E_I$ . The dimension of our latent space is 620, which is the same as our word embedding dimension. In implementation, we fix  $E_I$  and  $E_T$  and append projection matrix  $\theta_{P_I}$  and  $\theta_{P_T}$  as the last layers. During training, we only update projection matrix. We initialize all weight matrix by uniform initialization. The number of hidden units used in LN-LSTM is 1000, and we optimize our model by Adam optimizer (Kingma and Ba 2015) with start learning rate as 0.001 and decayed factor as 0.5 every 80. We follow NST to use gradient clipping = 5 in DLN. The training epoch is 100 with batch size as 64. We choose  $\lambda = 0.5$  for DLN training.

**DLN-Multi** The vocabulary size of each domain is 5500. We choose  $\lambda_1 = 0.2$  and  $\lambda_2 = 0.1$ . Other settings and hyperparameters are the same as DLN except the regularization term

$$R = \|\theta'_W - \theta_W\|_2 + \|\theta'_V - \theta_V\|_2 + \|\theta'_{E_I} - \theta_{E_I}\|_2 \quad (20)$$

We implement the  $\theta'_W$  by  $\theta'_W = \theta_W \|\theta_{\Delta W}$ , where  $\|\cdot\|$  is the concatenation operation of matrix and  $\theta_{\Delta W}$  is the new vocabulary used in the new style. In the subtraction, we only subtract the  $\theta_W$  part in  $\theta'_W$  to match the dimension of matrix. Similar logic can be applied to  $\theta'_V$ . We use the projection weight  $\theta_{P_I}$  learned during pre-training and the weight during the training of DLN-Mutli as our  $\theta'_{E_I}$  and  $\theta_{E_I}$ .

**StyleNet** To train StyleNet, we use the same hidden unit of LSTM as DLN and follow the iterative training method reported in the original paper except that we also update the share weight when training language model on target corpus, which we found this modification has better convergence in our task as shown in Fig 8. We also apply our decay learning setting, which we found has faster convergence. During inference, we follow StyleNet and NST by using beam search with a beam width of 5 and the unknown token for all methods.

**Pretrained classifier in transfer accuracy.** For the classifier used in evaluation, we use convolutional network proposed in (Kim 2014). We train the classifier to achieve over 99% accuracy to distinguish  $\mathcal{D}_S$  and  $\mathcal{D}_T$ .

### Human Evaluation Setup

We performed two human evaluation tasks using the Amazon Mechanical Turk<sup>5</sup> platform. The first was a relevance task, asking how well does descriptions relate to the image content on a four level scale. We provide screen-shots of the instructions given to workers in Fig 9. The second study aims to compare the attractiveness of descriptions generated by the DLN mode and DLN-Multi. Fig 9 is the screen-shots given to works for this experiment. To ensure reliable results and avoid workers who choose randomly, only workers with more than 70% accuracy and 100 successful HITs previously are allowed to attend the study. In the second human study, we also provide a dummy text as trap option to monitor the labelling quality. The result shows almost no trap options are chosen, indicating experiment results to be reliable.

### Qualitative example of stylish image description generation

We demonstrate more qualitative examples of stylish image description generated by DLN in Fig 11.

<sup>4</sup><https://keras.io/applications/>

<sup>5</sup><https://www.mturk.com>

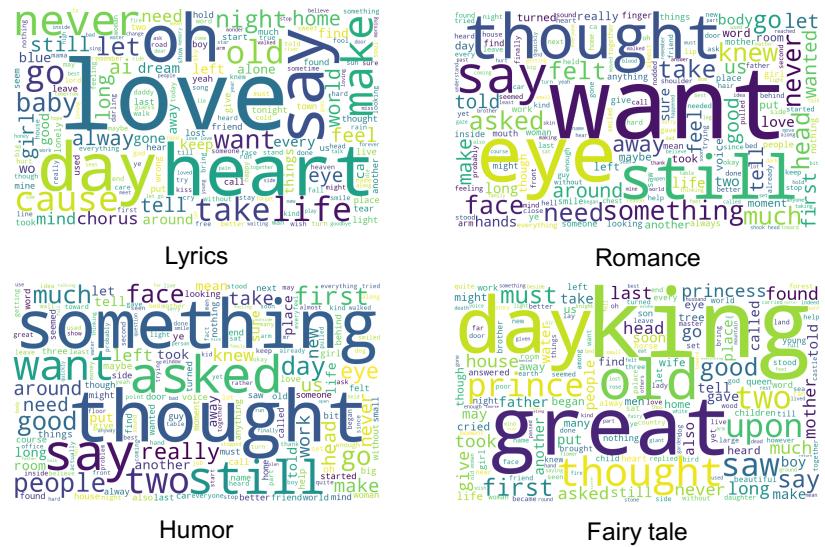


Figure 7: Word cloud of our four datasets

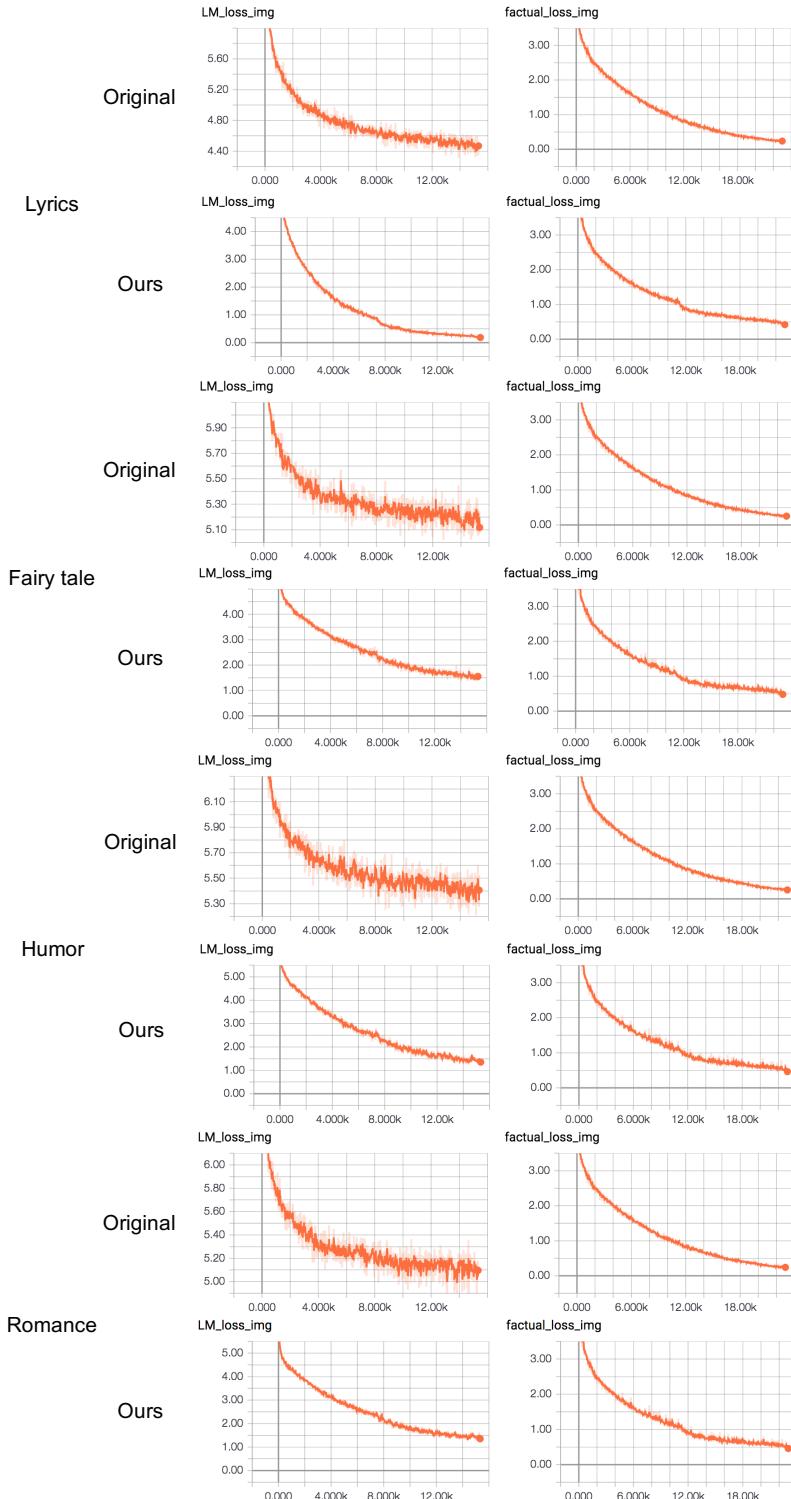


Figure 8: Comparison of training StyleNet by method mentioned in original paper (Original) and our modification (Ours). The LM\\_loss and factual\\_img\\_loss refer to  $\mathcal{L}_T$  and  $\mathcal{L}_S$  respectively

## Please evaluate the relevance between the image and paragraph below

### Instructions

You are given an image and a paragraph from **novels**. Please evaluate how well does the paragraph content relate to objects that appear in the image. For example, if there are houses and dogs in the image, words like dog, house, or **other semantic related nouns**(Ex. animal, cat, building, apartment, home) are considered relating to the image content.

Meaning of each options are explained below.

- **Related:** Most nouns in the paragraph are related to the image content.
- **Partly related:** About half nouns in the paragraph are related to image content.
- **Weakly related:** A few nouns in the paragraph are related to image content
- **Unrelated:** None of nouns in the paragraph are related to image content.

### Image to be described:



### Paragraph:

This is exactly what he was playing on his tennis court. He felt an overwhelming compassion for him, and when you hit him.

Related

Partly related

Weakly related

Unrelated

Figure 9: A screen-shot of human study on semantic relevance of generated paragraphs.

## Please choose the most attractive paragraph

### Instructions

Paragraphs below should describe the given image with a specific style (e.g., tell a romantic story based on wedding ceremony picture). Please choose the most attractive paragraph based on the following **PRIORITIES**:

1. **Content:** How well does the paragraph describe **objects appeared in the image**. For example, if the dog and house appear in the image, choose the paragraph which contains more words like dog, house, or other semantic related words(Ex. animal, cat, building, apartment)
2. **Style:** How well does the paragraph **match the specific style**.
3. **Naturalness:** How similar the paragraph is to human language.**More fluent and more grammatically correct** paragraph is better.

Therefore, you should first consider the content. If it's difficult to choose one option among all candidates, you should consider style then consider naturalness.

### Target style:

Fairy tale

### Image to be described:



- Sed ut perspiciatis unde omnis iste
- The clock tower made very much like Mary, and hardly had taken so much of it that there would be soon after all the fairies began to be most anxious for all that.
- There's a church at the market, which, from the house, the rooms of the cities; and the other time, two and people, every day is done;

Figure 10: A screen-shot of human study on paragraph attractiveness.

## Fairy Tale



**GT (unstylish):** This image is at a train station. There is a train on the tracks. The train is white, red and blue. There are people waiting on the train. The train tracks are clean.

**DLN:** This train is stopped on the train tracks filled with fear, and as soon as the train passed over a river, it was ready to start again.

## Romance



**GT (unstylish):** A white eagle flying over the water by a dock. The sky is cloudy and dark. There are ships docked at the port. There are people with yellow raincoat sitting on top of the boat. The bird wings are spread out and the bird is about to make a landing attempt.

**DLN:** A very dark and still day we were at war. I know why you take care of the water, but I shoot every drop. They're here, being on the water. I only ask for once. Now that starts my morning, my bird is floating on its surface.

## Humor



**GT (unstylish):** A black cat is on the brown desk. The cat is on the border of the desk. A glass with black drink is on back the cat. Behind the cat there is a laptop open with a blue screen. On the screen of the laptop there is icons color white on the right. Next to the wall there is a black box and a tan horizontal box.

**DLN:** A cat is looking back at me. My cat is sitting on the computer desk in the black chair, which is near the door of my computer and TV. He has no idea what he and sign in return for a computer, but His head under the desk too. What do you think of him?

## Lyrics



**GT (Unstylish):** There is a red light on a pole hanging over the street. There are buildings lining the street. There are cars driving on the street. There is a light post next to the street.

**DLN:** A busy town's of a road an road life town tall; But I can see my old grey look out of your face; Look at me now, You're still had inside of me; [Chorus]; You should know where I'm going; You got to keep a country train; And just to see your face in the desert sand; And the train on earth would you know; To welcome me home; Back home to the old home; Back home to the old home

Figure 11: Examples of stylistic image description by DLN.