StyleNet: Generating Attractive Visual Captions with Styles

Chuang Gan[1] Zhe Gan[2] Xiaodong He[3] Jianfeng Gao [3] Li Deng[3]
[1] IIIS, Tsinghua University, China
[2] Duke University, USA
[3] Microsoft Research Redmond, USA

Abstract

We propose a novel framework named StyleNet to address the task of generating attractive captions for images and videos with different styles. To this end, we devise a novel model component, named factored LSTM, which automatically distills the style factors in the monolingual text corpus. Then at runtime, we can explicitly control the style in the caption generation process so as to produce attractive visual captions with the desired style. Our approach achieves this goal by leveraging two sets of data: 1) factual image/video-caption paired data, and 2) stylized monolingual text data (e.g., romantic and humorous sentences). We show experimentally that StyleNet outperforms existing approaches for generating visual captions with different styles, measured in both automatic and human evaluation metrics on the newly collected FlickrStyle10K image caption dataset, which contains 10K Flickr images with corresponding humorous and romantic captions.

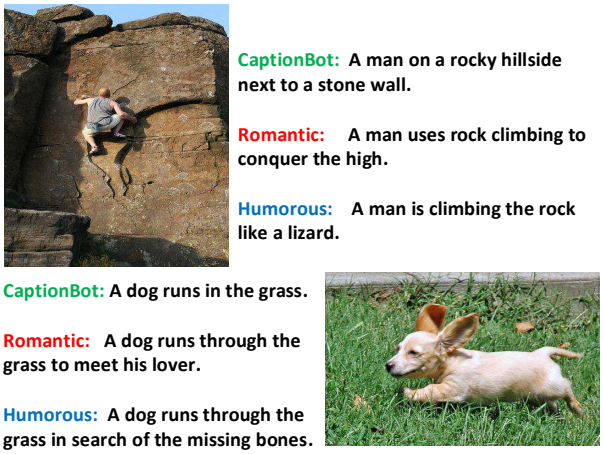

Figure 1. We address the problem of visual captioning with Styles. Given an image, our StyleNet can generate attractive image captions with different styles.

1. Introduction

Generating a natural language description of an image is an emerging interdisciplinary problem at the intersection of computer vision, natural language processing, and artificial intelligence. This task is often referred to as image captioning. It serves as the foundation of many important applications, such as semantic image search, visual intelligence in chatting robots, photo and video sharing on social media, and aid for people to perceive the world around them. However, we observed that the captions generated by most of the existing state-of-the-art image captioning systems [50, 32, 22, 5, 10, 9, 52, 54, 55, 2, 46] usually provide a factual description of the image content, while style is the often-overlooked element in the caption generation process. These systems usually use a language generation model that mixes the style with other linguistic patterns of language generation, thereby lacking a mechanism to control the style explicitly.

On the other hand, a stylized (*e.g.*, romantic or humorous) description will greatly enrich the expressibility of the caption and make it more attractive. An attractive image caption will add more visual interest to images and can even become a distinguishing trademark of the system. This is particularly valuable for certain applications such as increasing user engagement in chatting bots, or enlightening users in photo captioning for social media.

Figure 1 gives two examples to illustrate the setting of the problem. For the image at the top, the Microsoft CaptionBot [46] produces a caption that reads "*A man on a rocky hillside next to a stone wall*". Compared to this factual caption, the proposed StyleNet is able to generate captions with specific styles. For example, if romantic style is required, it describes the image as "*A man uses rock climbing to conquer the high*", while the caption is "*A man is climbing the mountain likes a lizard*" if a humorous style is demanded. Similarly, for the image at the bottom, the Microsoft CaptionBot produces a caption like "*A dog runs in the grass*". In contrast, the StyleNet can describe this image in a romantic style, such as "*A dog runs through the grass to*

meet his lover", or in a humorous style, "*A dog runs through the grass in search of the missing bones*". Compared to the flat description that most current captioning systems have produced, the stylized captions not only are more expressive and attractive, but also make images become more popular and memorable. The task of image captioning with styles will also facilitate many real-world applications. For example, people enjoy sharing their photos on social media, such as Facebook, Flickr, etc. However, users always struggle to come up with an attractive title when uploading them. Therefore, it is valuable if the machine could automatically recommend attractive captions based on the content of the image.

Prior to our work, Alexander *et al.* [34] has investigated generating image captions with positive or negative sentiments, where sentiments could be considered as a kind of style. In order to incorporate sentiments into captions, they proposed a switching Recurrent Neural Network (RNN). Training the switching RNN requires not only paired image-sentiment caption data, but also word-level supervision to emphasize the sentiment words (*e.g.*, sentiment strengths of each word in the sentiment caption), which makes the approach very expensive and difficult to scale up.

To address these challenges, we propose in this paper a novel framework, named as StyleNet, which is able to produce attractive visual captions with styles only using monolingual stylized language corpus (*i.e.* without paired images) and standard factual image/video-caption pairs. StyleNet is built upon the recently developed methods that combine Convolutional Neural Networks (CNNs) with Recurrent Neural Networks (RNNs) for image captioning. Our work is also motivated by the spirit of multitask sequence-to-sequence training [31]. Particularly, we introduce a novel factored LSTM model that can be used to disentangle the factual and style factors from the sentences through multi-task training. Then at running time, the style factors can be explicitly incorporated to generate different stylized captions for an image. We evaluate StyleNet on a newly collected Flickr stylized image caption dataset. Our results show that the proposed StyleNet significantly outperforms previous state-of-the-art image captioning approaches, measured by a set of automatic metrics and human evaluation. In summary, our work has made the following contributions:

- To the best of our knowledge, we are the first to investigate the problem of generating attractive image captions with styles without using supervised style-specific image-caption paired data.
- We propose an end-to-end trainable StyleNet framework, which automatically distills the style factors from monolingual textual corpora. In caption generation, the style factor can be explicitly incorporated to produce attractive captions with the desired style.
- We have collected a new Flickr stylized image caption dataset. We expect that this dataset can help advance the research of image captioning with styles.
- We demonstrate that our StyleNet framework and Flickr stylized image caption dataset can also be used to produce attractive video captions.

The rest of this paper is organized as follows. In Section 2, we review related work in image captioning. Section 3 presents the factored LSTM, a key building block of the proposed StyleNet framework. We show how the factored LSTM is applied to generate attractive image captions with different styles. We introduce the new collected Flickr stylized image caption dataset, called FlickrStyle10K, in Section 4. Experimental settings and evaluation results are presented in Section 5. Section 6 concludes the paper.

2. Related Work

Our paper relates mainly to two research topics: image captioning and unsupervised/semi-supervised captioning, which will be briefly reviewed in this section.

2.1. Image Captioning

Early approaches on image captioning could be roughly divided into two families. The first one is based on template matching [11, 27, 53, 29, 35]. These approaches start from detecting object, action, scene and attributes in images and then fill them into a hand-designed and rigid sentence template. The captions generated by these approaches are not always fluent and expressive. The second one is retrieval-based approaches. These approaches first retrieve the visually similarity images from a large database, and then transfer the captions of retrieved images to fit the query image [28, 36, 20, 41]. There is little flexibility to modify words based on the content of the query image, since they directly rely on captions of training images and could not generate new captions.

Recent successes of using neural networks in image classification [26, 43, 40, 17] , object detection [16, 15, 39] and attribute learning [12] motivates strong interests in using neural networks for image captioning [50, 32, 22, 5, 21, 10, 9, 52, 54, 55, 2, 46]. The leading neural-network-based approaches for automatic image captioning fall into two broad categories. The first one is the encoder-decoder based framework adopted from neural machine translation [42]. For instance, [50] extracted global image features using hidden activations of a CNN and then fed them into a LSTM which is trained to generate a sequence of words. [52] took one step further by introducing the attention mechanism, which selectively attends to different areas of the image when generating words one by one. [55] further improved the image captioning results by selectively attending to a set of semantic concepts extracted from the image to generate image captions. [54] introduced a reviewer module

to achieve attention mechanism. [51, 25] have investigated to generate dense image captions for individual regions in images. The other category of work is based on a compositional approach [10, 46]. For example, [10] employs a CNN to detect a set of semantic tags , then uses a maximum entropy language model to generate a set of caption candidates, and finally adopts a deep multimodal similarity model to re-rank the candidates to generate the final caption. Most recently, Gan *et al.* [13] proposed a novel semantic compositional network that extends each weight matrix of the LSTM to an ensemble of tag-dependent weight matrices and achieved state-of-the-art results on image captioning.

However, despite encouraging progress on generating fluent and accurate captions, most image captioning systems only produce factual descriptions of the images, including people, objects, activity, and their relations. The styles that make image caption attractive and compelling have been mostly neglected. [34] proposed to generate positive and negative sentiment captions with a switching RNN model, which is relevant to our work. [14] investigated to generate descriptive caption for visually impaired people. However, our work differs from them in two respects. First, we focus our study on generating humorous and romantic captions, aid for making image captions attractive and compelling for applications on social media. Second, our proposed StyleNet only takes the external language corpus as supervision without paired images, which are much cheaper than the word-level supervision used in the switching RNN model, thus more suitable to scale up.

2.2. Semi-supervised/Unsupervised Captioning

Our work is also relevant to semi-supervised and unsupervised visual captioning. [48] investigated the use of distributional semantic embeddings and LSTM-based language models trained on external text corpora to improve visual captioning. [38] proposed to use a variational autoencoder to improve captioning. [31] proposed a multitask sequence-to-sequence-learning framework to improve image captioning by joint training using external text data for other tasks. However, they have not explored how to distill the style factors learned from external text data to generate attractive image captions with styles. In more recent work, Mao *et al.* [33] and Hendricks *et al.* [18] proposed to generate descriptions for objects unseen in paired training data by learning to transfer knowledge from seen objects. Different from transferring the relationships between seen and unseen object categories, we propose StyleNet to separate the style factor from the generic linguistic patterns in caption generation so as to transfer the styles learned from monolingual text data for attractive visual captioning.

3. Approach

In this section, we describe our method of generating attractive image captions with styles. We first briefly review the LSTM model and how it is applied to image captioning [50]. We then introduce the factored LSTM module, which serves as the building block of StyleNet. Finally, we will describe StyleNet, which is end-to-end trained by leveraging image-caption paired data and additional monolingual language corpus with certain styles. The framework of our StyleNet is illustrated in Figure 2.

3.1. Caption Generation with LSTM

The Long Short-term Memory (LSTM) [19] model is a special type of RNNs that solves the vanishing and exploding gradients problem of conventional RNN architectures. The core of the LSTM architecture is the memory cell, which encodes the knowledge of the input at every time step that has been observed and the gates which determine when and how much the information conveys. Particularly, there are three gates: the input gate $\boldsymbol{i}_t$ to control the current input $\boldsymbol{x}_t$, the forget gate $\boldsymbol{f}_t$ to forget previous memory $\boldsymbol{c}_{t-1}$, and the output gate $\boldsymbol{o}_t$ to control how much of the memory to transfer to the hidden state $\boldsymbol{h}_t$. Together, they enable the LSTM to model long-term dependencies in sequential data. The gates and the cell updating rules in time t in an LSTM block is defined as follows:

$$\boldsymbol{i}_t = \text{sigmoid}(\boldsymbol{W}_{ix}\boldsymbol{x}_t + \boldsymbol{W}_{ih}\boldsymbol{h}_{t-1}) \quad (1)$$

$$\boldsymbol{f}_t = \text{sigmoid}(\boldsymbol{W}_{fx}\boldsymbol{x}_t + \boldsymbol{W}_{fh}\boldsymbol{h}_{t-1}) \quad (2)$$

$$\boldsymbol{o}_t = \text{sigmoid}(\boldsymbol{W}_{ox}\boldsymbol{x}_t + \boldsymbol{W}_{oh}\boldsymbol{h}_{t-1}) \quad (3)$$

$$\tilde{\boldsymbol{c}}_t = \tanh(\boldsymbol{W}_{cx}\boldsymbol{x}_t + \boldsymbol{W}_{ch}\boldsymbol{h}_{t-1}) \quad (4)$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \tilde{\boldsymbol{c}}_t \quad (5)$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \odot \boldsymbol{c}_t \quad (6)$$

$$\boldsymbol{p}_{t+1} = \text{Softmax}(\boldsymbol{C}\boldsymbol{h}_t) \quad (7)$$

where $\odot$ denotes element-wise product. The hidden state $\boldsymbol{h}_t$ is then fed into a Softmax to produce the probability distribution over all words in the vocabulary. The variable $\boldsymbol{x}_t$ is the element of the input sequence at time step t, and $\boldsymbol{W}$ denoted the LSTM parameters to be learned. Specifically, $\boldsymbol{W}_{ix}$, $\boldsymbol{W}_{fx}$, $\boldsymbol{W}_{ox}$, and $\boldsymbol{W}_{cx}$ are the weight matrices applied to the input variable $\boldsymbol{x}_t$, and $\boldsymbol{W}_{ih}$, $\boldsymbol{W}_{fh}$, $\boldsymbol{W}_{oh}$, and $\boldsymbol{W}_{ch}$ are the weight matrices applied to recurrently update the values of hidden states.

The recipe for caption generation with the CNN and RNN models follows the encoder-decoder framework originally used in neural machine translation [42, 6, 1], where an encoder is used to map the sequence of words in the source language into a fixed-length vector, and a decoder, once initialized by that vector, is used to generate the words in the target language one by one. During training, the goal is

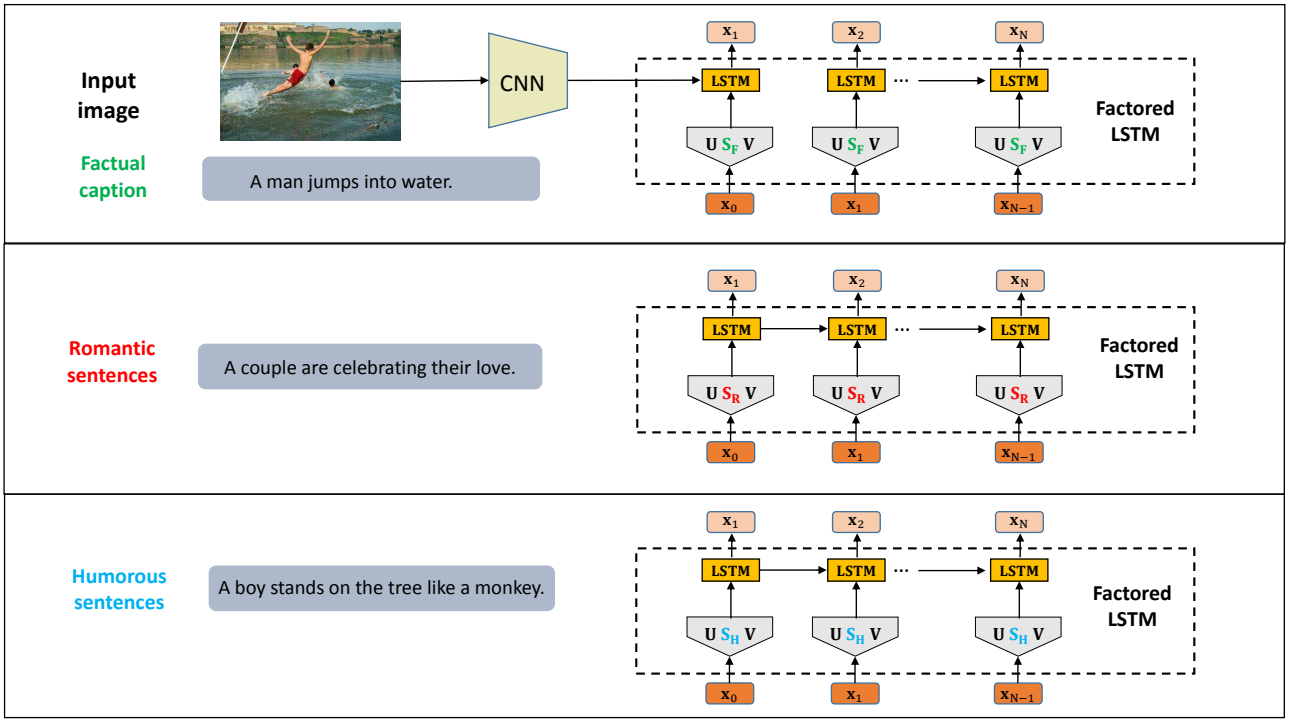

Figure 2. The framework of the StyleNet. We illustrate learning the StyleNet using the image and factual-caption paired data, plus monolingual romantic-style and humorous-style text corpora. During training, the factored LSTM-based decoders, which share the same set of parameters except the style specific factor matrix (*e.g.*, $\boldsymbol{S}_F$ for the factual style, $\boldsymbol{S}_R$ for the romantic style, and $\boldsymbol{S}_H$ for the humorous style, respectively), are trained on these data via multi-task learning.

to minimize the total cross-entropy loss given the source-target sentence pairs. When applying this framework to image caption generation, the task can be considered as translating from images to the target language. The commonly used strategies in literature [50, 52, 32] are to adopt a pre-trained CNN model as an encoder to map an image to a fixed dimensional feature vector and then use a LSTM model as the decoder to generate captions based on the image vector.

3.2. Factored LSTM Module

In this section, we describe a variant of the LSTM model, named as Factored LSTM, which serves as a major building block of StyleNet. The traditional LSTM used in image captioning mainly captures long-term sequential dependencies among the words in the sentences, but fails to factor the style from other linguistic patterns in the language. To remedy this issue, we propose a Factored LSTM module, which factors the parameters $\boldsymbol{W}_x$ in the traditional LSTM model into three matrices $\boldsymbol{U}_x$, $\boldsymbol{S}_x$, $\boldsymbol{V}_x$, as follows:

$$\boldsymbol{W}_x = \boldsymbol{U}_x \boldsymbol{S}_x \boldsymbol{V}_x \tag{8}$$

Suppose $\boldsymbol{W}_x \in \mathbb{R}^{\mathsf{M}\times\mathsf{N}}$, then $\boldsymbol{U}_x \in \mathbb{R}^{\mathsf{M}\times\mathsf{E}}$, $\boldsymbol{S}_x \in \mathbb{R}^{\mathsf{E}\times\mathsf{E}}$ and $\boldsymbol{V}_x \in \mathbb{R}^{\mathsf{E}\times\mathsf{N}}$. We apply this factored module to the input weight matrices including $\boldsymbol{W}_{ix}$, $\boldsymbol{W}_{fx}$, $\boldsymbol{W}_{ox}$, and $\boldsymbol{W}_{cx}$ that are used to transform the input variable $\boldsymbol{x}_t$, which fuels the content of the caption and influence the style directly. We leave the recurrent weight matrices, including $\boldsymbol{W}_{ih}$, $\boldsymbol{W}_{fh}$, $\boldsymbol{W}_{oh}$, and $\boldsymbol{W}_{ch}$, which mainly capture the long span syntactic dependency of the language, unchanged. Accordingly, the memory cells and gates in the proposed Factored LSTM are defined as follows:

$$\boldsymbol{i}_t = \text{sigmoid}(\boldsymbol{U}_{ix}\boldsymbol{S}_{ix}\boldsymbol{V}_{ix}\boldsymbol{x}_t + \boldsymbol{W}_{ih}\boldsymbol{h}_{t-1}) \tag{9}$$

$$\boldsymbol{f}_t = \text{sigmoid}(\boldsymbol{U}_{fx}\boldsymbol{S}_{fx}\boldsymbol{V}_{fx}\boldsymbol{x}_t + \boldsymbol{W}_{fh}\boldsymbol{h}_{t-1}) \tag{10}$$

$$\boldsymbol{o}_t = \text{sigmoid}(\boldsymbol{U}_{ox}\boldsymbol{S}_{ox}\boldsymbol{V}_{ox}\boldsymbol{x}_t + \boldsymbol{W}_{oh}\boldsymbol{h}_{t-1}) \tag{11}$$

$$\tilde{\boldsymbol{c}}_t = \tanh(\boldsymbol{U}_{cx}\boldsymbol{S}_{cx}\boldsymbol{V}_{cx}\boldsymbol{x}_t + \boldsymbol{W}_{ch}\boldsymbol{h}_{t-1}) \tag{12}$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \tilde{\boldsymbol{c}}_t \tag{13}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \odot \boldsymbol{c}_t \tag{14}$$

$$\boldsymbol{p}_{t+1} = \text{Softmax}(\boldsymbol{C}\boldsymbol{h}_t) \tag{15}$$

In the factored LSTM model, the matrix sets $\{\boldsymbol{U}\}$, $\{\boldsymbol{V}\}$, and $\{\boldsymbol{W}\}$ are shared among different styles, which are designed to model the generic factual description inside all the text data. The matrix set $\{\boldsymbol{S}\}$, however, is style specific, thus to distill underlying style factors in the text data. Specifically, we denote $\boldsymbol{S}_F$ the set of factor matrices for the factual style in the standard language description, $\boldsymbol{S}_R$ the set of factor matrices for the style of romantic, and $\boldsymbol{S}_H$ the

set of factor matrices for the style of humorous.

3.3. Training StyleNet

In order to learn to disentangle the style factors from the text corpus, we use an approach similar to multi-task sequence to sequence training [31].

There are two kinds of tasks the factored LSTM model needs to optimize. In the first task, the factored LSTM is trained to generate factual captions given the paired images. In the second task, the factored LSTM is trained as a language model. Note that the parameters of the factored LSTMs for both tasks are shared, except the style-specific factor matrices. Therefore, according to this design, the shared parameters model the generic language generation process, while the style-specific factor matrix captures the unique style of each stylized language corpus. The loss function across different tasks is the negative log likelihood of word $\boldsymbol{x}_t$ at each time step t.

As shown in figure 2, in training the LSTM will start with an initial state transformed from a visual vector when trained with a paired image, and start with a random noise vector otherwise. More specifically, for the first task that needs to train a factored LSTM model using the image and factual captions paired data, we first encode the image into a fixed-length vector, *i.e.*, a single feature vector obtained by extracting the activation of a pre-trained CNN, and then we map it via a linear transformation matrix $\boldsymbol{A}$ to an embedding space for initializing the LSTM. For the language side, each word is firstly represented as a one-hot vector and is then mapped to a continuous space via an word embedding matrix $\boldsymbol{B}$. During training, we only feed the visual input to the first step of the LSTM, following [50]. The parameters of LSTM to be updated in training include the linear transformation matrix $\boldsymbol{A}$ for transforming image features, the word embedding matrix $\boldsymbol{B}$, and the parameters of factored LSTM, including the shared matrix sets $\{\boldsymbol{U}\}$, $\{\boldsymbol{V}\}$, $\{\boldsymbol{W}\}$, and the factual-style specific matrix set $\boldsymbol{S}_F$.

Then we also need to train the factored LSTM to capture the stylized language patterns. During our multi-task training, in the second task, the factored LSTM is trained as a language model on the romantic sentences or humorous sentences. The word embedding matrix $\boldsymbol{B}$ and the parameters $\{\boldsymbol{U}\}$, $\{\boldsymbol{V}\}$, $\{\boldsymbol{W}\}$ are also shared across data with different styles. However, we will only update either the romantic-style specific matrix set $\boldsymbol{S}_R$ or the humorous-style specific matrix set $\boldsymbol{S}_H$, when trained on the romantic or humorous sentences, respectively. Since the matrix set $\{\boldsymbol{S}\}$ is style specific while all other parameters of the LSTM are shared across all tasks, the model is imposed to use $\{\boldsymbol{S}\}$ to distill the unique style factors contained in each language corpus, and other parameters to model the general language generation process.

At running time, we use style-specific factor matrix $\boldsymbol{S}$ plus other shared parameter set to form a factored LSTM according to equations (9)- (15). Then we extract and transform the feature vector of a given image, and feed it into the factored-LSTM based decoder to produce the caption with the desired style.

4. Creating Flickr Stylized Caption Dataset

To facilitate research in stylized image captioning, we have collected a new dataset called FlickrStyle10K, which is built on Flickr 30K image caption dataset [20]. We present the details of this dataset in the rest of this section.

4.1. Data Collection

Inspired by previous work [4, 56, 20], we have used the Amazons Mechanical Turk to gather caption annotations. However, collecting both accurate and attractive image captions with styles is much more challenging than collecting traditional visual captions. It took quite some iterations to test and evaluate user interfaces and instructions for collecting stylized captions. For example, we first instructed the annotator to directly write one humorous and one romantic caption given an image. However, we found it is difficult to control the quality of the captions written under this instruction. The annotators often wrote some phrases or comments that are irrelevant to the content of the image. Such kind of data is hardly useful for facilitating research on modeling the style factors in visual captioning.

Therefore, instead of asking the annotators to directly write a new caption, we switch the task to editing image captions. We showed a standard factual caption for an image, and then asked the annotators to revise the caption to make it romantic or humorous. We also gave some examples of factual captions and the corresponding humorous or romantic modifications. In practice, we have observed that the captions under these instructions are both relevant to image content and capture the required style sufficiently.

4.2. Quality Control

To ensure the quality of the collected Stylized image caption dataset, we first only allow workers who have completed at least 500 previous HITs with 90% accuracy to access our annotations. We also include some additional reviewers to check the quality of the resulting captions through Amazon Mechanical Turk. Three workers were assigned per stylized image caption and each worker was asked to rank whether it has the desired style. And we only maintain the images captions that have more than two hits.

In total, our Flickr stylized image caption dataset, called FlickrStyle10K, contains 10K images. We split the data into 7K for training, 2K for validation and 1K for testing, respectively. For the training and validation sets, we collect one humorous caption and one romantic caption for each image. For testing set, we collect five humorous and romantic

Romantic References							
Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	CIDEr	METEOR
CaptionBot [46]	40.4	20.2	12.7	7.6	0.36	0.26	0.133
NIC [50]	42.0	21.4	12.5	7.8	0.36	0.28	0.134
Fine-tuned	43.2	21.6	12.7	7.6	0.34	0.24	0.139
Multi-task [31]	44.1	23.7	14.3	9.5	0.36	0.29	0.145
StyleNet (F)	41.2	21.4	12.1	7.7	0.36	0.24	0.135
StyleNet (R)	**46.1**	**24.8**	**15.2**	**10.4**	**0.38**	**0.31**	**0.154**
Humorous References							
Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	CIDEr	METEOR
CaptionBot [46]	43.4	21.4	12.2	7.1	0.35	0.21	0.134
NIC [50]	43.1	22.8	13.2	7.9	0.36	0.23	0.136
Fine-tuned	43.0	20.7	12.9	7.8	0.34	0.19	0.128
Multi-task [31]	47.1	23.9	13.9	8.8	0.37	0.25	0.148
StyleNet (F)	42.9	22.3	12.9	7.7	0.36	0.23	0.135
StyleNet (H)	**48.7**	**25.4**	**14.6**	**10.1**	**0.38**	**0.27**	**0.152**

Table 1. Compared image caption results with baseline approaches on the FlickrStyle10K dataset.

captions written by five independent AMT worker for evaluation. In addition to the newly collected stylized captions, each image in this dataset also has 5 factual captions, as provided from the Flikcr30K data set [20].

5. Experiments

To validate the effectiveness of StyleNet, we have conducted experiments on both image and video captioning.

5.1. Experiments on Image Captioning

5.1.1 Experimental Setup

Dataset We first evaluate StyleNet on the newly collected FlickrStyle10K dataset which contains ten thousand Flickr images with stylized captions. We used the 7K images with their factual captions to train the factual image captioning model. For the additional text corpus, we used the 7K stylized captions without paired images to train the stylized language model.

Images and captions pre-processing We extract the 2,048-dimensional feature vector from the last pooling layer of the ResNet152 model [17], which is pre-trained on ImageNet dataset [7], for each image, and then transform it into a 300-dimensional vector as the visual input for captioning. For the captions, we first construct a word vocabulary which consists of the words occurring more than 2 times in the factual caption and maintain all the words occurred in the stylized captions. Each word in the sentence is represented as a one-hot vector, which has a value of one only in the element corresponding to the index of the word, and zeros in other elements. Then we transform this one-hot word vector to a 300-dimensional vector through a word embedding matrix.

Evaluation Metric To evaluate the captions generated by StyleNet, we used four metrics which are commonly used in image captioning, including BLEU [37], METEOR [8], ROUGE [30] and CIDEr [47]. For all four metrics, a larger score means better performance. We further conduct human evaluation through the Amazons Mechanical Turk. We ask the judgers to select the most attractive captions given the image in the potential scenarios of image sharing on social media.

Compared Baselines To evaluate the performance of the proposed StyleNet in generating attractive image captions with styles, we compared with four strong baseline approaches, namely:

- **Neural Image Caption (NIC)** [50]: we implement NIC with a standard LSTM and the encoder-decoder image caption pipeline. We train it by using the factual image-caption pairs of the FlickrStyle10K dataset.
- **CaptionBot** [46]: the commercial image captioning system released by Microsoft, which is trained on the large-scale factual image-caption pair data.
- **Multi-task** [31]: we implement a traditional LSTM in the multitask sequence learning framework as presented in [31].
- **Fine-tuned**: we first train an image caption model using the factual image-caption paired data in FlickrStyle10K, and then use the additional stylized text data to update the parameters of the LSTM language model.

Implementation Details We implement the StyleNet using Theano [44]. Both caption and language models are trained using Adam [24] algorithms. We set the batch size for image captioning model and stylized language models as 64 and 96, respectively; the learning rate is set to 0.0002

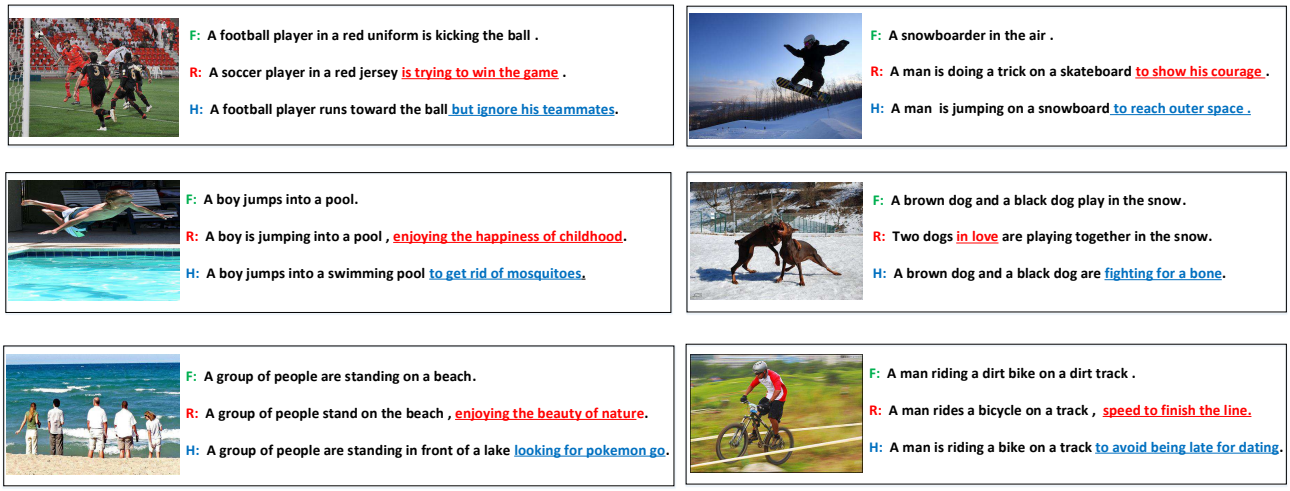

Figure 3. Examples of different style captions generated by the StyleNet.

NIC	CaptionBot	StyleNet (R)	StyleNet (H)
6.4%	7.8%	45.2%	40.6%

Table 2. Human voting results for the attractiveness of generated image captions.

and 0.0005, respectively. We set the units of LSTM cell and the factored matrix as 512. All the parameters are initialized by a uniform distribution. For multi-task training, we adopt the alternating training approach, where each task is optimized for one epoch, and then switched to the next task. We start with the image captioning task and then transfer to the stylized language modelling task. We try to combine the romantic and humorous style together in training, but do not observe further improvements. The training will converge in 30 epochs. Given test images, we generate the captions by performing a beam-search with a beam size of 5.

For comparsion, we use the same visual features extracted from ResNet 152 for the StyleNet and all other baselines (except CaptionBot). We train the NIC model by setting the batch size as 64 and terminate the training after 20 epochs based on the performance on the validation set. For the CaptionBot baseline, we directly use the captions generated by the Microsoft Computer Vision API which powers the CaptionBot [46].

We use the same visual features and vocabulary as in StyleNet for the fine-tuned and muti-task baselines. For the fine-tuned model, we first trained an image captioning model for 20 epochs by setting the learning rate as 0.0002, and then trained the stylized language model for 25 epochs by setting learning rate as 0.0005. For the multitask baseline reported in Table 1, it is implemented using the same setting as the StyleNet, but replace the factored LSTM model to a traditional LSTM model. All the parameters except the image feature transformation matrix $\boldsymbol{A}$ are shared among different tasks. We observed that the performance started to converge after 30 epochs.

5.1.2 Experimental Results

We summarized the experimental results in Table 1. The notations of StyleNet (F), StyleNet (R), and StyleNet (H) denote the standard factual captioning, romantic style captioning, and humorous style captioning using StyleNet, respectively. The name of other baselines in Table 1 is self-explained. In evaluation, we report the results using both the romantic references and the humorous references. From Table 1, we observe that, (1) given a desired style, the StyleNet that tailors to that style achieves the best results over the baseline approaches across multiple automatic evaluation metric; (2) the StyleNet can effectively model the style factors in caption generation, as demonstrated by the relative performance variance. For example, the StyleNet equipped with the correct style factor matrices gives superior performances, while other StyleNet variants perform comparable to baselines when the quality of the captions is measured against the corresponding stylized references (romantic and humorous) as ground truth; (3) the proposed Factored LSTM outperforms models based on traditional LSTM across different metrics, showing the effectiveness of the factored LSTM for distilling the style from language corpus.

We also report the human evaluation results in Table 2. For each image, we present four captions generated by NIC, CaptionBot, StyleNet with a romantic style, and StyleNet with a humorous style in a random order to judges, and ask them to select the most attractive captions, considering the scenario of sharing images with captions on social media. The results in Table 2 indicate that nearly 85% of the judges think the captions generated by StyleNet, either in a roman-

Video	StyleNet (H)	StyleNet (R)
17.2%	39.1%	43.7%

Table 3. Human voting results for the attractive of video captions.

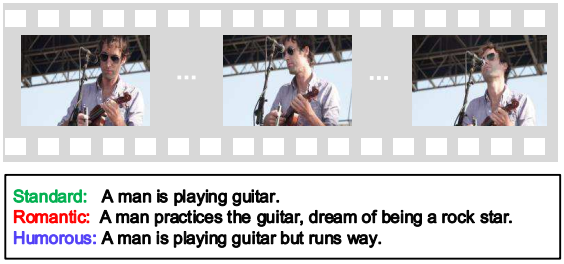

Figure 4. Examples of different style video captions generated by the StyleNet.

tic style or a humorous style, are more attractive than factual captions from traditional captioning systems.

We further investigate the output of the StyleNet, and present some typical examples in Figure 3. We can see that the captions with the standard factual style only describe the facts in the image in a dull language, while both the romantic and humorous style captions not only describe the content of the image, but also express the content in a romantic or humorous way through generating phrases that bear a romantic (*e.g.* in love, the happiness of childhood, the beauty of nature, win the game, etc) or humorous (*e.g.* get rid of mosquitoes, reach outer space, pokemon go, bone, etc) sense. More interestingly, besides being humorous or romantic, the phrases that the StyleNet generates fit the visual content of the image coherently, making the caption visually relevant and attractive.

5.2. Experiments on Video Captioning

To further evaluate the versatility of the proposed StyleNet framework, we extend the StyleNet to the video captioning task by using the FlickrStyle10K dataset and the videos-caption paired data in the Youtube2text dataset [3].

5.2.1 Experimental Setup

Youtube2Text is a commonly used dataset for research in video captioning, which contains 1,970 Youtube clips, and each clip is annotated with about 40 captions. We follow the standard split defined by [49], *i.e.*, 1,200 videos for training, 100 videos for validation, and 670 videos for testing. We use the 3D CNN (C3D) [45] pre-trained on the Sport 1M dataset [23] to construct video-clip features from both spatial and temporal dimensions. We then use average pooling to obtain the video-level representations, which is a fixed-dimension vector. We use that video-level feature vector as the visual input to StyleNet. At the language side, we preprocess the descriptions the same way as that for the image captioning task. We further transform the video feature vector and text one-hot vectors into 300-dimensional space through two different transformation matrices. The hyper parameters of factored LSTM and the training mechanism are the same as in the image captioning task. The training converges after 30 epochs. We compared the StyleNet with the baseline, called *Video*, which is a standard video captioning model using video-caption paired data.

5.2.2 Experimental Results

We report the experimental results in Table 3, which shows the human preference of the video captioning generated by the baselines and StyleNet. For each video clip, we generate three captions using the Video baseline and the humorous and romantic style captions by StyleNet, respectively. We then show the video clip and the captions to AMT judges in a random order, and ask them to select the most attractive caption by sharing video clips and captions on social media. Similar to the observation in image captioning experiments, we find that over 80% of judges favor the captions generated by StyleNet with either a romantic or a humorous style. Compared to the baseline trained on the video data, the StyleNet can learn the style factor from the stylized monolingual text corpus, plus learn from the video-caption data to capture the factual part during the video caption generation, demonstrating great versatility. We present several caption examples from StyleNet in Figure 4. We observed that the StyleNet can effectively control the style to generate both visually relevant and attractive captions for videos.

6. Conclusions

In this paper, we aim to generate attractive visual captions with different styles. To this end, we have developed an end-to-end trainable framework, named as StyleNet. By using a specialized factored LSTM module and through multi-task learning, StyleNet is able to learn styles from monolingual textual corpus. At running time, the style factor can be incorporated into the visual caption generation process through the factored LSTM module. Our quantitative and qualitative results demonstrate that the proposed StyleNet can indeed generate visually relevant and attractive captions with different styles. To facilitate future research on this emerging topic, we have collected a new Flickr stylized caption dataset, which will be released to the community.

Acknowledgement. Chuang Gan was partially supported by the National Basic Research Program of China Grant 2011CBA00300, 2011CBA00301, the National Natural Science Foundation of China Grant 61033001, 61361136003.

References

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *ICLR*, 2015. 3

[2] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *NIPS*, pages 1171–1179, 2015. 1, 2

[3] D. L. Chen and W. B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *ACL*, pages 190–200, 2011. 8

[4] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 5

[5] X. Chen and C. Lawrence Zitnick. Mind's eye: A recurrent visual representation for image caption generation. In *CVPR*, pages 2422–2431, 2015. 1, 2

[6] K. Cho, B. V. Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *EMNLP*, 2014. 3

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 6

[8] M. Denkowski and A. Lavie. Meteor universal: Language specific translation evaluation for any target language. In *ACL*, 2014. 6

[9] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, pages 2625–2634, 2015. 1, 2

[10] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, et al. From captions to visual concepts and back. In *CVPR*, pages 1473–1482, 2015. 1, 2, 3

[11] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *ECCV*, pages 15–29, 2010. 2

[12] C. Gan, T. Yang, and B. Gong. Learning attributes equals multi-source domain generalization. In *CVPR*, pages 87–97, 2016. 2

[13] Z. Gan, C. Gan, X. He, Y. Pu, K. Tran, J. Gao, L. Carin, and L. Deng. Semantic compositional networks for visual captioning. *CVPR*, 2017. 3

[14] S. Gella and M. Mitchell. Residual multiple instance learning for visually impaired image descriptions. *NIPS Women in Machine Learning Workshop*, 2016. 3

[15] R. Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015. 2

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *Computer Science*, pages 580–587, 2014. 2

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CVPR*, 2016. 2, 6

[18] L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell. Deep compositional captioning: Describing novel object categories without paired training data. *CVPR*, 2016. 3

[19] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. 3

[20] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013. 2, 5, 6

[21] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars. Guiding the long-short term memory model for image caption generation. In *ICCV*, pages 2407–2415, 2015. 2

[22] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015. 1, 2

[23] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. 8

[24] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *ICLR*, 2015. 6

[25] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *arXiv preprint arXiv:1602.07332*, 2016. 3

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *NIPS*, 2012. 2

[27] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Babytalk: understanding and generating simple image descriptions. In *CVPR*, pages 1601–1608, 2011. 2

[28] P. Kuznetsova, V. Ordonez, T. L. Berg, and Y. Choi. TREETALK: composition and compression of trees for image descriptions. *TACL*, 2:351–362, 2014. 2

[29] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi. Composing simple image descriptions using web-scale n-grams. In *ACL*, 2011. 2

[30] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8, 2004. 6

[31] M.-T. Luong, Q. V. Le, I. Sutskever, O. Vinyals, and L. Kaiser. Multi-task sequence to sequence learning. *ICLR*, 2015. 2, 3, 5, 6

[32] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-RNN). *ICLR*, 2015. 1, 2, 4

[33] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Learning like a child: Fast novel visual concept learning from sentence descriptions of images. In *ICCV*, 2015. 3

[34] A. Mathews, L. Xie, and X. He. Senticap: Generating image descriptions with sentiments. *AAAI*, 2015. 2, 3

[35] M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. Berg, K. Yamaguchi, T. Berg, K. Stratos, and I. Hal Daum?? Midge: generating image descriptions from computer vision detections. In *EACL*, pages 747–756, 2012. 2

[36] V. Ordonez, G. Kulkarni, T. L. Berg, V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. *NIPS*, pages 1143–1151, 2011. 2

[37] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318, 2002. 6

[38] Y. Pu, Z. Gan, R. Henao, X. Yuan, C. Li, A. Stevens, and L. Carin. Variational autoencoder for deep learning of images, labels and captions. In *NIPS*, pages 2352–2360, 2016. 3

[39] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2016. 2

[40] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *Computer Science*, 2014. 2

[41] C. Sun, C. Gan, and R. Nevatia. Automatic concept discovery from parallel text and visual corpora. In *ICCV*, pages 2596–2604, 2015. 2

[42] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *NIPS*, pages 3104–3112, 2014. 2, 3

[43] C. Szegedy, W. Liu, Y. Jia, and P. Sermanet. Going deeper with convolutions. *CVPR*, pages 1–9, 2015. 2

[44] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv: 1605.02688*, 2016. 6

[45] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3D convolutional networks. In *ICCV*, pages 4489–4497, 2015. 8

[46] K. Tran, X. He, L. Zhang, J. Sun, C. Carapcea, C. Thrasher, C. Buehler, and C. Sienkiewicz. Rich image captioning in the wild. *arXiv preprint arXiv:1603.09016*, 2016. 1, 2, 3, 6, 7

[47] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575, 2015. 6

[48] S. Venugopalan, L. A. Hendricks, R. Mooney, and K. Saenko. Improving lstm-based video description with linguistic knowledge mined from text. 2016. 3

[49] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko. Translating videos to natural language using deep recurrent neural networks. *NAACL*, 2015. 8

[50] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164, 2015. 1, 2, 3, 4, 5, 6

[51] L. Wei, Q. Huang, D. Ceylan, E. Vouga, and H. Li. Densecap: Fully convolutional localization networks for dense captioning. *Computer Science*, 2015. 3

[52] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057, 2015. 1, 2, 4

[53] Y. Yang, C. L. Teo, Daum, H. Iii, and Y. Aloimonos. Corpus-guided sentence generation of natural images. In *EMNLP*, pages 444–454, 2011. 2

[54] Z. Yang, Y. Yuan, Y. Wu, R. Salakhutdinov, and W. W. Cohen. Encode, review, and decode: Reviewer module for caption generation. *NIPS*, 2016. 1, 2

[55] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. *CVPR*, 2016. 1, 2

[56] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2014. 5