# Diverse Image Captioning with Grounded Style

Franz Klein[1], Shweta Mahajan[1(✉)], and Stefan Roth[1,2]

[1] Department of Computer Science, TU Darmstadt, Darmstadt, Germany
shweta.mahajan@visinf.tu-darmstadt.de
[2] hessian.AI, Darmstadt, Germany

**Abstract.** Stylized image captioning as presented in prior work aims to generate captions that reflect characteristics beyond a factual description of the scene composition, such as sentiments. Such prior work relies on *given* sentiment identifiers, which are used to express a certain global style in the caption, *e.g.* positive or negative, however without taking into account the stylistic content of the visual scene. To address this shortcoming, we first analyze the limitations of current stylized captioning datasets and propose COCO attribute-based augmentations to obtain varied stylized captions from COCO annotations. Furthermore, we encode the stylized information in the latent space of a Variational Autoencoder; specifically, we leverage extracted image attributes to explicitly structure its sequential latent space according to different localized style characteristics. Our experiments on the Senticap and COCO datasets show the ability of our approach to generate accurate captions with diversity in styles that are grounded in the image.

**Keywords:** Diverse image captioning · Stylized captioning · VAEs

## 1 Introduction

Recent advances in deep learning and the availability of multi-modal datasets at the intersection of vision and language [26,47] have led to the successful development of image captioning models [3,13,29,31,39]. Most of the available datasets for image captioning, *e.g.* COCO [26], consist of several ground-truth captions per image from different human annotators, each of which factually describes the scene composition. In general, captioning frameworks leveraging such datasets deterministically generate a single caption per image [5,11,21,23,24,28,32,45]. However, it is generally not possible to express the entire content of an image in a single, human-sounding sentence. Diverse image captioning aims to address this limitation with frameworks that are able to generate several *different* captions for a single image [4,30,43]. Nevertheless, these approaches

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/978-3-030-92659-5_27.

C. Bauckhage et al. (Eds.): DAGM GCPR 2021, LNCS 13024, pp. 421–436, 2021.
https://doi.org/10.1007/978-3-030-92659-5_27

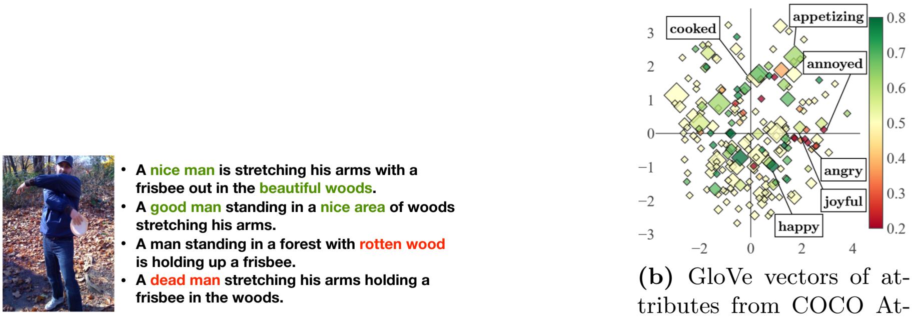
largely ignore image and text properties that go beyond reflecting the scene composition; in fact, most of the employed training datasets hardly consider such properties.

Stylized image captioning summarizes these properties under the term *style*, which includes variations in linguistic style through variations in language, choice of words and sentence structure, expressing different emotions about the visual scene, or by paying more attention to one or more localized concepts, *e.g.* attributes associated with objects in the image [35]. To fully understand and reproduce the information in an image, it is inevitable to consider these kinds of characteristics. Existing image captioning approaches [19,33,35] implement style as a global sentiment and strictly distinguish the sentiments into 'positive', 'negative', and sometimes 'neutral' categories. This simplification ignores characteristics of styles that are crucial for the comprehensive understanding and reproduction of visual scenes. Moreover, they are designed to produce one caption based on a given sentiment identifier related to one sentiment category, ignoring the actual stylistic content of the corresponding image [22,33,35].

In this work, we attempt *(1)* to obtain a more diverse representation of style, and *(2)* ground this style in attributes from localized image regions. We propose a Variational Autoencoder (VAE) based framework, Style-SeqCVAE, to generate stylized captions with styles expressed in the corresponding image. To this end, we address the lack of image-based style information in existing captioning datasets [23,33] by extending the ground-truth captions of the COCO dataset [23], which focus on the scene composition, with localized attribute information from different image regions of the visual scene [38]. This style information in the form of diverse attributes is encoded in the latent space of the Style-SeqCVAE. We perform extensive experiments to show that our approach can indeed generate captions with image-specific stylized information with high semantic accuracy *and* diversity in stylistic expressions.

## 2 Related Work

*Image Captioning.* A large proportion of image captioning models rely on Long Short-Term Memories (LSTMs) [20] as basis for language modeling in an encoder-decoder or compositional architecture [16,23,27,32,42,44]. These methods are designed to generate a single accurate caption and, therefore, cannot model the variations in stylized content for an image. Deep generative model-based architectures [4,13,30,31,43] aim to generate multiple captions, modeling feature variations in a low-dimensional space. Wang et al. [43] formulate constrained latent spaces based on object classes in a Conditional Variational Autoencoder (CVAE) framework. Aneja et al. [4] use a sequential latent space to model captions with Gaussian priors and Mahajan et al. [30] learn domain-specific variations of images and text in the latent space. All these approaches, however, do not allow to directly control the intended style to be reflected in each of the generated captions. This is crucial to generate captions with stylistic variation grounded in the images. In contrast, here we aim to generate diverse captions with the many localized styles representative of the image.



**(a)** Example from the Senticap dataset.

**(b)** GloVe vectors of attributes from COCO Attributes [38].

**Fig. 1.** *(a)* Senticap example ground-truth captions with globally positive or negative sentiment. *(b)* GloVe vectors of captions with adjectives from COCO Attributes. The color indicates the attribute SentiWordNet score [6] and the scale indicates the frequency of occurrence in the entire dataset. The original SentiWordNet scores are rescaled between 0 for the most negative and 1 for the most positive sentiment.

*Stylized Image Captioning.* Recent work has considered the task of stylized caption generation [9,33,40,46]. Mathews et al. [33] utilize 2000 stylized training captions in addition to those available from the COCO dataset to generate captions with either positive or negative sentiments. Shin et al. [40] incorporate an additional CNN, solely trained on weakly supervised sentiment annotations of a large image corpus scraped from different platforms. These approaches, however, rely on *given* style indicators during inference to generate new captions, to emphasize sentiment words in stylized captions, which may not reflect the true sentiment of the image. A series of works attempts to overcome the lack of available stylized captioning training data by separating the style components and the remaining information from the textual description such that they can be trained on both factual caption-image pairs and a distinct stylized text corpus. Gan et al. [17] share components in the LSTM over sentences of a particular style. Similarly, Chen et al. [9] use self-attention to adaptively consider semantics or style in each time step. You et al. [46] add a sentiment cell to the LSTM and Nezami et al. [35] apply the well established semantic attention idea to stylized captioning. Various approaches utilize adversarial formulations to generate stylized captions [19,22,36]. However, they also rely on given globally positive or negative sentiment identifiers to generate varied captions. Since an image can contain a variety of localized styles, in this work, we instead focus on generating diverse captions where the style information is *locally grounded in the image.*

## 3 Image Captioning Datasets with Stylized Captions

We begin by discussing the limitations of existing image captioning datasets with certain style information in the captions, specifically the Senticap dataset

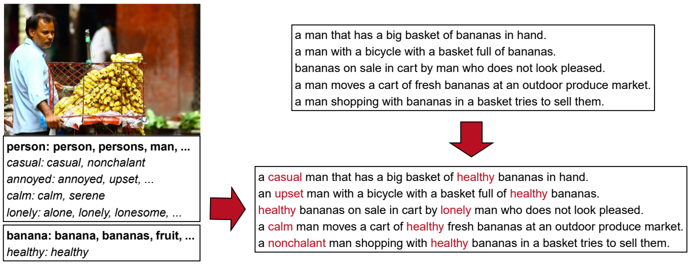
[33]. Further, we introduce an attribute (adjective) insertion method to extend the COCO dataset with captions containing different localized styles.

*Senticap.* Besides its limited size (1647 images and 4419 captions in the training set), the Senticap dataset [33] does not provide a comprehensive impression of the image content in combination with different sentiments anchored in the image (Fig. 1a). The main issue in the ground-truth captions is that the positive or negative sentiments may not be related to the sentiment actually expressed in the image. Some adjectives may even distort the image semantics. For instance, the man shown in Fig. 1a is anything but dead, though that is exactly what the last ground-truth caption describes. Another issue is the limited variety: There are 842 positive adjective-noun pairs (ANPs) composed of 98 different adjectives combined with one out of 270 different nouns. For the negative set only 468 ANPs exist, based on 117 adjectives and 173 objects. These adjectives, in turn, appear with very different frequencies in the caption set, *cf.* Fig. 3a.

*COCO Attributes.* The attributes of COCO Attributes [38] have the big advantage over Senticap that they actually reflect image information. Furthermore, the average number of 9 attribute annotations per object may reflect different possible perceptions of object characteristics. On the downside, COCO contains fairly neutral, rather positive than negatively connoted images: for instance, many image scenes involve animals, children, or food. This is reflected in the sentiment intensity of the associated attributes, visualized in Fig. 1b. Most of them tend to be neutral or positive; the negative ones are underrepresented.

### 3.1 Extending Ground-Truth Captions with Diverse Style Attributes

To address the lack of stylized ground-truth captions, we combine COCO captions [26], focusing on the scene composition, with style-expressive adjectives in COCO Attributes [38]. We remove 98 attribute categories that are less relevant for stylized captioning (*e.g.*, “cooked”) and define sets of synonyms for the remaining attributes to increase diversity. Some of the neutral adjective attributes are preserved since recognizing and dealing with object attributes having a neutral sentiment is also necessary to fully solve captioning with image-grounded styles. However, most of the 185 adjectives are either positive or negative. Likewise for the various COCO object categories, we initially define sets of nouns that appear interchangeably in the corresponding captions to name this object category. Subsequently, we iterate over all COCO images with available COCO Attributes annotations. Given a COCO image, associated object/attribute labels, and the existing ground-truth captions, we locate nouns in the captions that also occur in the sets of object categories to insert a sampled adjective of the corresponding attribute annotations in front of it. A part-of-speech tagger [8] helps to insert the adjective at the right position. This does not protect against an adjective being associated with the wrong noun if it occurs multiple times in the same caption. However, our observations suggest this


person: person, persons, man, ...
casual: casual, nonchalant
annoyed: annoyed, upset, ...
calm: calm, serene
lonely: alone, lonely, lonesome, ...
banana: banana, bananas, fruit, ...
healthy: healthy

a man that has a big basket of bananas in hand.
a man with a bicycle with a basket full of bananas.
bananas on sale in cart by man who does not look pleased.
a man moves a cart of fresh bananas at an outdoor produce market.
a man shopping with bananas in a basket tries to sell them.

a casual man that has a big basket of healthy bananas in hand.
an upset man with a bicycle with a basket full of healthy bananas.
healthy bananas on sale in cart by lonely man who does not look pleased.
a calm man moves a cart of healthy fresh bananas at an outdoor produce market.
a nonchalant man shopping with healthy bananas in a basket tries to sell them.

**Fig. 2.** Example of COCO caption augmentation by insertion of random samples from the COCO Attributes synonym sets in front of the nouns of COCO object categories.

to be rare. The example of the augmentation in Fig. 2 illustrates that inserting various adjectives could reflect a certain level of ambiguity in perception. Based on the COCO 2017 train split [23], this gives us a set of 266K unique, adjective-augmented and image-grounded training captions, which potentially reflect the image style.

### 3.2 Extending the Senticap Dataset

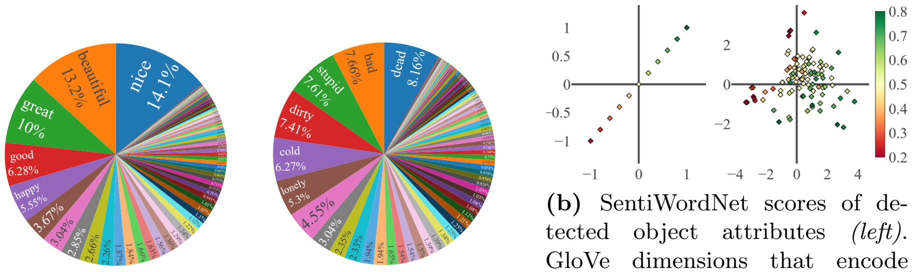
Since most prior work is evaluated on the Senticap dataset [22,33–35], whose characteristics strongly differ from COCO Attributes, using COCO Attributes-augmented captions for training and the Senticap test split for evaluation would result in scores that provide only little information about the actual model capabilities. We thus generate a second set of augmented COCO captions, based on the Senticap ANPs. These ANPs indicate which nouns and adjectives jointly appear in Senticap captions, independent of the sentiment actually expressed by the underlying image. Hence, to insert these sentiment adjectives into captions such that they represent the image content, we locate nouns appearing in the ANPs to sample and insert one of the adjectives corresponding to the detected nouns, utilizing the method from above. We thus obtain two different sets of training captions: COCO Attributes-augmented captions paired with COCO Attributes and Senticap-augmented captions, which are composed of 520K captions with positive adjectives and 485K captions with negative adjectives.

## 4 The Style-SeqCVAE Approach

To obtain image descriptions with styles grounded in images, we first extract the attributes associated with objects in the visual scene. Equipped with these style features, we formalize our Style-SeqCVAE with a structured latent space to encode the localized image-grounded style information.

### 4.1 Image Semantics and Style Information

Dense image features and corresponding object detections are extracted using a Faster R-CNN [18]. In order to obtain the different object attributes, we add a



**(a)** Frequency of positive *(left)* and negative *(right)* captions.

**(b)** SentiWordNet scores of detected object attributes *(left)*. GloVe dimensions that encode the most sentiment information *(right)*.

**Fig. 3.** *(a)* Adjective frequency in Senticap captions. *(b)* Two different latent space structuring approaches.

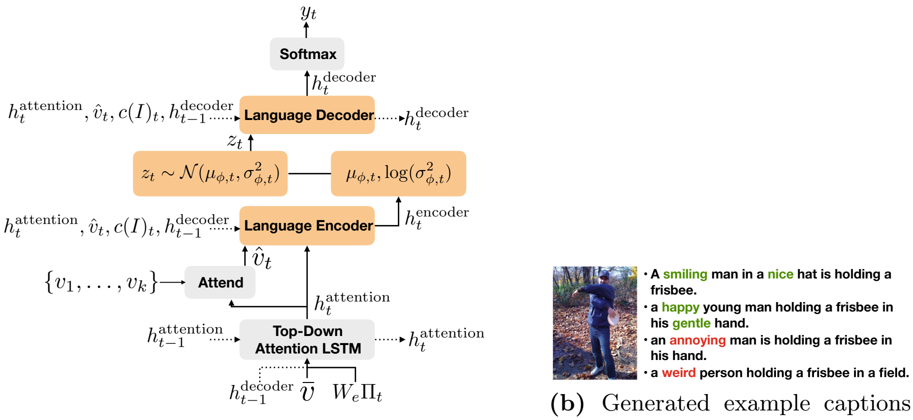
dense layer with sigmoid activation for multi-label classification. The loss term of the classification layer $L_{\text{CN}}$, which traditionally consists of a loss term for background/foreground class scores ($L_{\text{cls}}$), and a term for regression targets per anchor box ($L_{\text{reg}}$), is extended with another loss $L_{\text{att}}$:

$$\begin{aligned} L_{\text{CN}}(\{p_i\},\{b_i\},\{a_i\}) = \frac{1}{N_{\text{cls}}}\sum_i L_{\text{cls}}\left(p_i, p_i^*\right) + \lambda_1 \frac{1}{N_{\text{reg}}}\sum_i p_i^* L_{\text{reg}}\left(b_i, b_i^*\right) \\ + \lambda_2 \frac{1}{N_{\text{att}}}\sum_i \beta_i L_{\text{att}}\left(a_i, a_i^*\right), \end{aligned} \tag{1}$$

where $a_i^*$ are ground-truth attribute annotations and $a_i$ denotes the predicted probabilities for each particular attribute category being present at anchor $i$. $L_{\text{cls}}$ is the binary cross-entropy loss between the predicted probability $p_i$ of anchor $i$ being an object and the ground-truth label $p_i^*$. The regression loss $L_{\text{reg}}$ is the smooth $L_1$ loss [18] between the predicted bounding box coordinates $b_i$ and the ground-truth coordinates $b_i^*$. $L_{\text{att}}$ is the class-balanced softmax cross-entropy loss [12]. $N_{\text{cls}}$, $N_{\text{reg}}$, and $N_{\text{att}}$ are the normalization terms. $\lambda_1$ and $\lambda_2$ are the regularization parameters. Here, $\beta_i = 1$ if there is an attribute associated with anchor $i$ and $\beta_i = 0$ otherwise.

### 4.2 The Style-Sequential Conditional Variational Autoencoder

The goal of our Style-SeqCVAE framework is to generate diverse captions that reflect different perceivable style expressions in an image. We illustrate the proposed model in Fig. 4a. Consider an image $I$ and caption sequence $x = (x_1, \ldots, x_T)$, the visual features $\{v_1, \ldots, v_K\}$ for $K$ regions of the image are extracted from a Faster R-CNN (*cf.* Eq. 1) and the mean-pooled image features $\bar{v} = \frac{1}{K}\sum_k v_k$ are input to the attention LSTM [3]. In this work, we propose to further encode region-level style information in $c(I)_t$ (as discussed below), and update it at each time step using the attention weights ($\alpha_t$). This is based on the observation that the image styles can vary greatly across different


$y_t$
Softmax
$h_t^{\text{decoder}}$
$h_t^{\text{attention}}, \hat{v}_t, c(I)_t, h_{t-1}^{\text{decoder}}$
Language Decoder
$h_t^{\text{decoder}}$
$z_t$
$z_t \sim \mathcal{N}(\mu_{\phi,t}, \sigma^2_{\phi,t})$
$\mu_{\phi,t}, \log(\sigma^2_{\phi,t})$
$h_t^{\text{attention}}, \hat{v}_t, c(I)_t, h_{t-1}^{\text{decoder}}$
Language Encoder
$h_t^{\text{encoder}}$
$\hat{v}_t$
$\{v_1, \ldots, v_k\}$
Attend
$h_t^{\text{attention}}$
$h_{t-1}^{\text{attention}}$
Top-Down Attention LSTM
$h_t^{\text{attention}}$
$h_{t-1}^{\text{decoder}}$
$\bar{v}$
$W_e\Pi_t$

- A smiling man in a nice hat is holding a frisbee.
- a happy young man holding a frisbee in his gentle hand.
- an annoying man is holding a frisbee in his hand.
- a weird person holding a frisbee in a field.

**(a)** Style-Sequential CVAE architecture.

**(b)** Generated example captions with positive and negative styles.

**Fig. 4.** *(a)* Style-Sequential CVAE for stylized image captioning: overview of one time step. *(b)* Captions generated with Style-SeqCVAE on Senticap.

regions. To take this into account, we model a sequential VAE with explicit latent space structuring with an LSTM-based language encoder and language decoder (Fig. 4a highlights this in yellow). $h_t^{\text{attention}}$, $h_t^{\text{encoder}}$, and $h_t^{\text{decoder}}$ denote the hidden vectors of the respective LSTMs at a time step $t$. Encoding the latent vectors $z_t$ at each time step based on the reweighted image regions and corresponding component vectors $c(I)_t$ enables the model to structure the latent space at the image region level instead of only globally. Similar to Anderson et al. [3], the relevance of the input features at a specific time step $t$ depends on the generated word $W_e\Pi_t$, where $W_e$ is a word embedding matrix and $\Pi_t$ is the one-hot encoding of the input word. Given the attended image feature $\hat{v}_t$, the latent vectors $z$ are encoded in the hidden states $h_{t-1}^{\text{attention}}$ and $h_{t-1}^{\text{decoder}}$. Moreover, to allow for image-specific localized style information, we enforce an attribute-based structured latent space.

The log-evidence lower bound at time step $t$ is given by

$$\begin{aligned}\log p\left(x_t|I, x_{<t}, z_{\leq t}, c(I)_t\right) \geq \mathbb{E}_{q_\phi}&\left[\log p_\theta\left(x_t|I, x_{<t}, z_{\leq t}, c(I)_t\right)\right]\\ &- D_{KL}[q_\phi(z_t|I, x_{<t}, z_{<t}, c(I)_t) \,\|\, p_\theta(z_t|c(I)_t)].\end{aligned} \tag{2}$$

Here, $p_\theta$ is the prior distribution parameterized by $\theta$ and $q_\phi$ denotes the variational posterior distribution with parameters $\phi$.

*Attribute-Specific Latent Space Structuring.* The choice of the prior contributes significantly to how the latent space is structured. The additive Gaussian prior has proven to be beneficial for both diversity and controllability of the caption generation process [4,43]. Unlike [43], where the latent space is constrained based on the objects, we instead leverage attributes to encode the styles in the image. Specifically, available attributes are explicitly assigned to one of the image features $\{v_1, ..., v_k\}$ and can thus be divided into subsets $A_k = \{a_{k,1}, ..., a_{k,j}\}$ containing $J_k$ different attributes. Furthermore, we adopt the attention mechanism of [2], which provides a set of weights $\alpha_t = \{\alpha_{t,1}, ..., \alpha_{t,k}\}$ to readjust the impact of each image feature $v_k$ at every time step $t$. Accordingly, a weight $\alpha_{t,k}$ is also mapped to the subset $A_k$, which belongs to the image feature $v_k$. This property is exploited for the attribute-specific latent space structuring and makes it possible to weight the contribution of each individual attribute set $A_k$. We assume that the approximate style of an image region $v_t$ is represented by the total of all associated attributes $A_k$. The additive Gaussian latent space can then be reformulated to calculate a $\mu_t$ at each time step $t$ as

$$\mu_t = \sum_{k=1}^{K} \frac{\alpha_{t,k}}{J_k} \sum_{i=1}^{J_k} \mu_{i,k}. \tag{3}$$

The prior mean $\mu_t$ is thus composed of the attention-weighted average of each image region-specific linear combination of $\mu_{i,k}$. The variance $\sigma_t^2$ does not need to be calculated explicitly as $\sigma_i^2$ is equal for all attribute categories $i$. However, randomly initialized, attribute category-specific Gaussian components $\mu_i$ do not reflect any semantic or contextual similarities between different attributes. Therefore, in this work two alternative attribute category-specific $\mu_i$ initialization approaches are pursued: In the first case, we uniformly initialize each $\mu_i$ to the SentiWordNet score [6] corresponding to attribute category $i$. In the following, this latent space is referred to as the *SentiWordNet* latent space. In the second case, both sentimental and semantic characteristics of different attributes are taken into account by initializing each $\mu_i$ with the $n$ dimensions of GloVe vectors corresponding to the attribute category $i$ that most strongly encode its word sentiments. These dimensions are identified by an application of PCA on the 20 COCO Attributes labels with the strongest SentiWordNet scores. If the number of extracted dimensions is less than the desired dimensionality $z$, its elements are simply repeated to upscale it. We refer to this setup as *SentiGloVe* latent space. Both latent space structuring approaches are exemplified in Fig. 3b. With $\mu_i$ already encoding attribute-specific information, it serves as explicit input $c(I)_t$ to the encoder and decoder, thus $c(I)_t = \mu_t$. While training, the encoder produces Gaussian parameters $\mu_{\phi,t}$ and $\log \sigma^2_{\phi,t}$, which are used to encode a latent $z_t$ and calculate the KL-Divergence

$$\begin{aligned} D_{KL}[q_\phi(z_t|I, x_{<t}, z_{<t}, c(I)_t) \,\|\, p(z_t|c(I)_t)] &= \log\left(\tfrac{\sigma_t}{\sigma_{\phi,t}}\right) + \tfrac{1}{2\sigma_t^2}\mathbb{E}_{q_\phi}\left[\|z_t - \mu_t\|^2\right] - \tfrac{1}{2} \\ &= \log\left(\tfrac{\sigma_t}{\sigma_{\phi,t}}\right) + \tfrac{\sigma^2_{\phi,t} + \|\mu_{\phi,t} - \mu_t\|^2}{2\sigma_t^2} - \tfrac{1}{2}, \end{aligned} \tag{4}$$

which is used to maximize the variational lower bound. Furthermore, $z_t$ is provided to the decoder to produce the output word $y_t$ of the current time step. During generation, the encoder is dropped and $z_t$ values are sampled from the same attribute-specific additive Gaussian prior that is used while training. The attributes attached to the image features are actual (hard) detections from the image feature extractor instead of ground-truth annotations.

*Sentiment-Specific Latent Space Structuring.* Furthermore, we extend our approach to the Senticap dataset available for stylized image captioning and thus allow for a quantitative comparison of the proposed framework with existing work. The Senticap dataset provides positive as well as negative captions for images in the dataset. This implies that the captions in the Senticap dataset do not represent a variety of image-anchored styles and therefore, we cannot expect the latent space structure defined above to be particularly helpful. Thus, we take into account the COCO captions which are augmented with Senticap adjectives and are either labeled as "positive" or "negative" and define a simple latent space structure around the two clusters. Additionally, a third cluster is defined for captions with a neutral sentiment, *e.g.* original COCO captions. In contrast to the attention-weighted, attribute-based latent space structures defined above, the mean remains constant over all time steps and only depends on the provided sentiment identifier in this setup. And since each caption is distinctively assigned to one of these three clusters, the additive Gaussian structuring is not suitable. Instead, the prior mean is fully defined by one of the three predefined clusters with values $c(I)_t \in \{-0.5, 0.0, 0.5\}$ depending on the negative, neutral, or positive sentiment identifier. In this case, $\mu_t = c(I)_t$ and $p(z_t|c(I)_t) = \mathcal{N}\left(z_t \;\middle|\mu_t, \sigma^2 I\right)$. Similar to the latent space structuring approaches described above, the prior variance $\sigma_t^2$ is initially set and identical for all three clusters. Having obtained these prior parameters and $c(I)_t$, the training procedure is identical to that of the attribute-specific latent space. During evaluation, we generate diverse captions for a given sentiment by selecting one of the three clusters (by choosing the prior mean associated with the intended sentiment).

*Style-Based Constrained Beam Search.* Since only a fraction of COCO images are annotated with attributes, we rely on COCO Attributes-augmented captions for a fraction of training images. When the model is now trained using a combination of the original and attribute-augmented COCO captions, the attribute words occur rarely in the decoding procedure, especially the ones that are extremely underrepresented in the training data. This issue is optionally addressed by presenting the detected object attributes as constraints during the decoding procedure using constrained beam search (CBS) [1].

## 5 Experiments

We next evaluate and analyze our approach for the generation of a diverse set of image captions for a particular image with image-grounded styles. We evaluate on the Senticap and COCO datasets. On Senticap, we use the Senticap-augmented captions for training. When evaluating on the standard COCO dataset, we utilize the COCO Attributes-augmented captions for training in order to encode image-specific style information in the latent space.

*Evaluation Metrics.* The accuracy of the captions is evaluated with standard metrics – Bleu (B) 1–4 [37], CIDEr (C) [10], ROUGE (R) [25], and METEOR

**Table 1.** Top-1 oracle scores on Senticap test split captions with positive and negative sentiments. $n$ denotes the number of captions generated per image.

| | | Positive | | | | | | | Negative | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | $n$ | B1 | B2 | B3 | B4 | R | C | M | B1 | B2 | B3 | B4 | R | C | M |
| Senticap [33] | 1 | 49.1 | 29.1 | 17.5 | 10.8 | 36.5 | 54.4 | 16.8 | 50.0 | 31.2 | 20.3 | 13.1 | 37.9 | 61.8 | 16.8 |
| StyleNet [17] | 1 | 45.3 | – | 12.1 | – | – | 36.3 | 12.1 | 43.7 | – | 10.6 | – | – | 36.6 | 10.9 |
| You et al. [46] | 1 | 51.2 | 31.4 | 19.4 | 12.3 | 38.6 | 61.1 | 17.2 | 52.2 | 33.6 | 22.2 | 14.8 | 39.8 | 70.1 | 17.1 |
| Chen et al. [9] | 1 | 50.5 | 30.8 | 19.1 | 12.1 | 38.0 | 60.0 | 16.6 | 50.3 | 31.0 | 20.1 | 13.3 | 38.0 | 59.7 | 16.2 |
| Senti-Attend [35] | 1 | 57.6 | 34.2 | 20.5 | 12.7 | 45.1 | 68.6 | 18.9 | 58.6 | 35.4 | 22.3 | 14.7 | 45.7 | 71.9 | 19.0 |
| MSCap [19] | 1 | 46.9 | – | 16.2 | – | – | 55.3 | 16.8 | 45.5 | – | 15.4 | – | – | 51.6 | 16.2 |
| AttendGAN [36] | 1 | 56.9 | 33.6 | 20.3 | 12.5 | 44.3 | 61.6 | 18.8 | 56.2 | 34.1 | 21.3 | 13.6 | 44.6 | 64.1 | 17.9 |
| Memcap [48] | 1 | 51.1 | – | 17.0 | – | – | 52.8 | 16.6 | 49.2 | – | 18.1 | – | – | 59.4 | 15.7 |
| Karayil et al. [22] | 1 | 54.7 | 34.6 | 22.0 | 14.4 | 41.8 | 46.1 | 18.5 | 57.0 | 36.2 | 23.4 | 15.1 | 44.5 | 50.9 | 19.9 |
| | 10 | 65.6 | 43.9 | 29.5 | 20.2 | 48.8 | 63.1 | 22.1 | **67.6** | 46.3 | 31.9 | 21.9 | 50.4 | 68.8 | 23.5 |
| Style-SeqCVAE | 1 | 53.8 | 33.2 | 20.1 | 12.5 | 41.5 | 71.1 | 19.7 | 55.2 | 34.5 | 21.5 | 13.4 | 41.5 | 75.5 | 19.4 |
| | 10 | **66.3** | **46.3** | **32.2** | **22.2** | **51.2** | **110.5** | **25.2** | 66.1 | **46.8** | **33.3** | **23.7** | **50.9** | **111.9** | **24.4** |

(M) [7]. Similar to Mathews et al. [33], we consider the percentage of candidate captions containing at least one of the Senticap ANPs as part of the evaluation, referred to as SEN%. Additionally, we report the precision (SP) and recall (SR) of sentiment adjectives occurring in candidate and reference captions.

### 5.1 Evaluation on the Senticap Dataset

In this setting, the original COCO 2017 train split is combined with the COCO captions augmented with Senticap adjectives. The available captions express either a negative, neutral, or positive sentiment and, therefore, the latent space is explicitly structured around three different clusters encoding the sentiment expressed in the generated captions. Unless otherwise stated, constrained beam search is not used for caption generation. Since the Senticap dataset consists of positive and negative ground-truth captions for images, not related to the actual image sentiment, prior work [9,17,19,22,33,35,36,46] generates a positive as well as a negative caption for a given image based on the style indicator. Therefore, in order to compare our approach on the Senticap evaluation dataset, we generate positive and negative captions for a given image based on the sentiment-specific latent space as discussed above (Sect. 4.2). The quality of the generated positive and negative captions is reported in Table 1. When generating only one caption per image ($n = 1$), the achieved scores are comparable to the best-performing existing work. When generating $n = 10$ captions, the presented approach performs clearly better than the only related work [22] that generates diverse captions conditioned on the style indicator. This implies that unlike our Style-SeqCVAE approach, [22] does not encode as many variations in style content for a given image. The fact that we obtain high scores with our approach on metrics that take longer $n$-grams into account indicates that appropriate adjectives related to style are inserted into the captions in a

**Table 2.** Additional evaluation on the Senticap test set. SP denotes the sentiment precision and SR the sentiment recall.

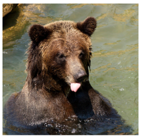
| n | Method | B1 | B2 | B3 | B4 | R | C | M | %SEN | SP | SR |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Senticap, Mathews et al. [33] | 48.8 | 29.8 | 18.7 | 11.8 | 37.2 | 56.6 | 16.8 | 87.5 | **0.33** | 0.14 |
|  | COCO + Senticap-augm. | 54.5 | 33.9 | 20.8 | 13.0 | 41.5 | 73.3 | 19.6 | 93.6 | 0.30 | 0.15 |
|  | COCO + Senticap-augm. + CBS | 54.7 | 34.0 | 21.1 | 13.2 | 41.7 | 74.1 | 19.7 | 100.0 | 0.30 | 0.15 |
| 10 | COCO + Senticap-augm. | 66.2 | 46.6 | 32.8 | 23.0 | 51.0 | 111.2 | 24.8 | 99.3 | 0.26 | 0.30 |
|  | COCO + Senticap-augm. + CBS | **66.3** | **47.2** | **33.6** | **23.9** | **51.5** | **114.5** | **25.3** | **100.0** | 0.25 | **0.32** |

suitable place. This also suggests that our proposed caption augmentation approach preserves the syntactic correctness of the resulting training captions. Most striking is the significant performance increase in the CIDEr score, which particularly rewards the use of $n$-grams that only rarely appear in the reference captions. This implies that explicitly structuring the latent space around fixed style-based clusters encourages even underrepresented style adjectives to prevail during decoding, especially when multiple captions are generated for an image.

In Table 2, we show that the diversity in the generated captions is the result of the different ways of expressing sentiment (and not solely based on diversity from the factual descriptions). In this setting, we include the Senticap training set without decoding constraints (COCO + Senticap-augm) and with decoding constraints (COCO + Senticap-augm + CBS). The proportion of reference sentiment adjectives appearing in the candidate captions doubles from ∼0.15 to ∼0.3 when producing 10 captions per image, which shows clearly that the diversity in the captions also has an effect on sentiment expression. Captions generated by [33] exclusively consider the most dominant adjectives in the training/test captions without much diversity. Thus, its slightly higher SP score is expected, given the massive adjective imbalance of Senticap. The Senticap-augmented model (COCO + Senticap-augm) inserts at least one ANP with matching sentiment in almost every caption (>93.6%). Furthermore, we do not observe a significant improvement in any metric when attribute-based CBS constraints are applied. This shows that the latent space of our Style-SeqCVAE can effectively model the style information in the latent space. Additionally, the high CIDEr score supports that the captions generated by our approach are accurate. Qualitative examples in Fig. 4b show the captions with varied styles generated with Style-SeqCVAE. For the given image, the diverse captions reflect the positive as well as negative sentiments showing that the latent space of our approach effectively captures style information of the data distribution.

### 5.2 Evaluation on the COCO Dataset

We now show that our approach along with the extended COCO Attributes-augmented captions can generate diverse captions with styles anchored in the image. Since the COCO test split contains only descriptions of the scene composition, it is difficult to quantitatively evaluate for stylized caption generation on the COCO dataset. Therefore, we first present a qualitative analysis of the



['angry', 'alone / lonely', 'furry', 'hairy', 'bulky', 'fluffy', 'fuzzy','playful', 'wild', 'natural', 'soft', 'strong', 'dangerous', 'heavy']

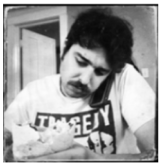
weak constraining

- a furry large brown bear sitting in the water.
- a close shot of a furry bear in the water with a fish.
- a furry young bear is swimming in the water.

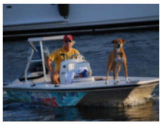
individual constraining

- a playful adult bear in a body of water.
- a dangerous bear that is sitting in the water.
- a furry bear swimming alone in water.

**Fig. 5.** Examples of captions generated by either constraining the decoding procedure on the whole set of detected image attributes (weak constraining) vs. one specific attribute as constraint per caption (individual constraining).



- black and white photo of a clean man talking on a cell phone.
- a black and white photo of a lonely man talking on a cell.
- a busy man with a cell phone in his hand.
- a man with a soft face talking on a cell phone.



- a serious man and his tame dog sit on a useful small boat in the water.
- a clean man and his tame dog on a simple boat in the ocean.
- a nonchalant man on a useful boat with his tame dog.
- a serious man on a useful boat with a tame dog on the front.

**Fig. 6.** Generated example captions with SentiGloVe-structured latent space and individual decoding constraints for images with multiple objects.

Style-SeqCVAE approach. For this, we consider captions generated directly from the latent space (without CBS) as well as with CBS constraints to account for the rarity of the attributes in the dataset. Unlike [1], which presents class labels as constraints, in this work, we enforce style information from automatically extracted image attributes as constraint to the caption generator. For an effective application of CBS constraints in conjunction with Style-SeqCVAE, two different decoding strategies are considered: In weak constraining, the constraint is to use at least one attribute from the set of detected attributes during decoding for a given image. The detected attributes are obtained directly from the image extractor (Faster R-CNN) trained using Eq. 1. In individual constraining, the attribute to be inserted is explicitly selected from the set of detected attributes for an image and provided as a constraint. Since the model is forced to take this attribute into account, it enables to also consider attributes that are underrepresented in the training data and which might otherwise be ignored. When generating multiple captions for a given image, a constraint is randomly selected from the set of detected attributes for each of the captions. This encourages diversity in the style of the generated captions for each image.

With the weak constraining mechanism, we observe that attributes that occur in high frequency in the training data are repeated in the generated captions. As shown in Fig. 5, the detected attribute "furry" associated with the object "bear" occurs across all the generated captions for the given image. In case of individual constraining, where an attribute is randomly provided as constraint from the set of detected attributes, this effect is not present. For example, in Fig. 5, we observe diversity in style with attributes like "playful" and "dangerous" describing the object "bear". In Fig. 6, we show an extension of the individual constraining procedure to occurrences of multiple objects in an image. Here, the model is forced to insert an attribute for at least two detected objects. For example, in

**Table 3.** Evaluation of semantic accuracy.

| Method | std | B1 | B2 | B3 | B4 | R | C | M |
|---|---|---|---|---|---|---|---|---|
| Div-BS [41] | – | 83.7 | 68.7 | 53.8 | 38.3 | 65.3 | 140.5 | 35.7 |
| AG-CVAE [43] | – | 83.4 | 69.8 | 57.3 | 47.1 | 63.8 | 125.9 | 30.9 |
| POS [14] | – | **87.4** | **73.7** | **59.3** | 44.9 | **67.8** | **146.8** | **36.5** |
| Seq-CVAE [4] | – | 87.0 | 72.7 | 59.1 | 44.5 | 67.1 | 144.8 | 35.6 |
| Style-SeqCVAE | 1 | 84.2 | 69.5 | 56.0 | 44.7 | 63.4 | 130.4 | 31.2 |
| | 2 | 86.6 | 72.0 | 58.8 | **47.6** | 65.9 | 137.2 | 32.8 |

**Table 4.** Caption diversity.

| Method | std | Div-1 | Div-2 |
|---|---|---|---|
| Div-BS [41] | – | 0.20 | 0.26 |
| AG-CVAE [43] | – | 0.24 | 0.34 |
| POS [14] | – | 0.24 | 0.35 |
| Seq-CVAE [4] | – | 0.25 | **0.54** |
| Style-SeqCVAE | 1 | 0.24 | 0.31 |
| | 2 | **0.29** | 0.43 |

Fig. 6 the diverse attributes are successfully inserted for the objects "man" and "face" or for "man", "dog", and "boat" in the captions of the respective images.

The quantitative evaluation of the proposed framework on the COCO dataset is limited due to the lack of ground-truth captions for direct comparison purposes (see supplemental). To obtain a better assessment in spite of that, we compare our method with various established approaches for diverse image captioning. Here, we use the SentiGloVe latent space to generate diverse captions. Following the standard evaluation protocol of [4,30,43], in Table 3 we show the top-1 oracle performance using 20 samples on various metrics for caption evaluation. We observe generally competitive performance, especially in the case where a standard deviation of 2 is used when sampling from the latent space. This is despite the fact that, unlike previous work on diverse image captioning, our model focuses on stylistic diversity, which is not represented in the ground-truth captions of the test set. The high accuracy scores, moreover, demonstrate the ability of the approach to successfully model the attribute-based style information to generate semantically coherent captions (*cf.* Fig. 5).

Furthermore, we quantitatively compare our approach against [4,14,41,43] for diversity. In Table 4, we use Div-1 and Div-2 for evaluation, where Div-$n$ is the ratio of distinct $n$-grams per caption to the total number of words generated per set of diverse captions. Following prior work, the scores are based on the top-5 captions with highest CIDEr scores of the COCO validation split [23] and consensus re-ranking [15] is applied before selecting the top-5 captions. Owing to the unconstrained latent space, Seq-CVAE [4] generates captions with high diversity. The scores on the diversity metrics indicate that in comparison to other approaches that also impose constraints in the latent space, *e.g.* AG-CVAE [43], our attribute-based latent space exhibits a higher diversity in style. This can be attributed to the structured sequential latent space, where the distribution of attributes is better captured conditioned on the image. This highlights the advantages of our style-specific latent space for diverse caption generation.

## 6 Conclusion

We present Style-SeqCVAE, a variational autoencoder framework to encode localized style information representative of the visual scene. The structured latent space exploits the object attributes from the associated images to model the characteristic styles. In particular, we leverage the attribute information in different image regions to express different styles present in the image. The key to the success of the proposed latent space is the combination of attribute-based style information and region-based image features via an attention mechanism for coherent caption generation. Our experiments demonstrate that our approach generates diverse and accurate captions with varied styles expressed in the image.

**Acknowledgement.** This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 866008).

## References

1. Anderson, P., Fernando, B., Johnson, M., Gould, S.: Guided open vocabulary image captioning with constrained beam search. In: EMNLP, pp. 936–945 (2017)
2. Anderson, P., Gould, S., Johnson, M.: Partially-supervised image captioning. In: NeurIPS, pp. 1875–1886 (2018)
3. Anderson, P., et al.: Bottom-up and top-down attention for image captioning and visual question answering. In: CVPR, pp. 6077–6086 (2018)
4. Aneja, J., Agrawal, H., Batra, D., Schwing, A.: Sequential latent spaces for modeling the intention during diverse image captioning. In: ICCV, pp. 4261–4270 (2019)
5. Aneja, J., Deshpande, A., Schwing, A.G.: Convolutional image captioning. In: CVPR, pp. 5561–5570 (2018)
6. Baccianella, S., Esuli, A., Sebastiani, F.: SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: LREC, pp. 2200–2204 (2010)
7. Banerjee, S., Lavie, A.: METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65–72 (2005)
8. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media, Inc. (2009)
9. Chen, C.K., Pan, Z., Liu, M.Y., Sun, M.: Unsupervised stylish image description generation via domain layer norm. In: AAAI, pp. 8151–8158 (2019)
10. Chen, X., et al.: Microsoft COCO captions: data collection and evaluation server. arXiv:1504.00325 (2015)
11. Chen, X., Zitnick, C.L.: Mind's eye: a recurrent visual representation for image caption generation. In: CVPR, pp. 2422–2431 (2015)
12. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: CVPR, pp. 9268–9277 (2019)
13. Dai, B., Fidler, S., Urtasun, R., Lin, D.: Towards diverse and natural image descriptions via a conditional GAN. In: ICCV, pp. 2970–2979 (2017)

14. Deshpande, A., Aneja, J., Wang, L., Schwing, A.G., Forsyth, D.: Fast, diverse and accurate image captioning guided by part-of-speech. In: CVPR, pp. 10695–10704 (2019)
15. Devlin, J., Gupta, S., Girshick, R., Mitchell, M., Zitnick, C.L.: Exploring nearest neighbor approaches for image captioning. arXiv:1505.04467 (2015)
16. Donahue, J., et al.: Long-term recurrent convolutional networks for visual recognition and description. TPAMI **39**(4), 677–691 (2017)
17. Gan, C., Gan, Z., He, X., Gao, J., Deng, L.: StyleNet: generating attractive visual captions with styles. In: CVPR, pp. 3137–3146 (2017)
18. Girshick, R.: Fast R-CNN. In: ICCV, pp. 1440–1448 (2015)
19. Guo, L., Liu, J., Yao, P., Li, J., Lu, H.: MSCap: multi-style image captioning with unpaired stylized text. In: CVPR, pp. 4204–4213 (2019)
20. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
21. Johnson, J., Karpathy, A., Fei-Fei, L.: DenseCap: fully convolutional localization networks for dense captioning. In: CVPR, pp. 4565–4574 (2016)
22. Karayil, T., Irfan, A., Raue, F., Hees, J., Dengel, A.: Conditional GANs for image captioning with sentiments. In: Tetko, I.V., Kůrková, V., Karpov, P., Theis, F. (eds.) ICANN 2019. LNCS, vol. 11730, pp. 300–312. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30490-4_25
23. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. TPAMI **39**(4), 664–676 (2017)
24. Kulkarni, G., et al.: BabyTalk: understanding and generating simple image descriptions. TPAMI **35**(12), 2891–2903 (2013)
25. Lin, C.Y.: ROUGE: a package for automatic evaluation of summaries. In: ACL Text Summarization Branches Out, pp. 74–81 (2004)
26. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
27. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: CVPR, pp. 3242–3250. IEEE Computer Society (2017)
28. Lu, J., Yang, J., Batra, D., Parikh, D.: Neural baby talk. In: CVPR, pp. 7219–7228 (2018)
29. Mahajan, S., Botschen, T., Gurevych, I., Roth, S.: Joint Wasserstein autoencoders for aligning multimodal embeddings. In: ICCVW, pp. 4561–4570 (2019)
30. Mahajan, S., Gurevych, I., Roth, S.: Latent normalizing flows for many-to-many cross-domain mappings. In: ICLR (2020)
31. Mahajan, S., Roth, S.: Diverse image captioning with context-object split latent spaces. In: NeurIPS, pp. 3613–3624 (2020)
32. Mao, J., Xu, W., Yang, Y., Wang, J., Yuille, A.L.: Deep captioning with multimodal recurrent neural networks (m-RNN). In: ICLR (2015)
33. Mathews, A., Xie, L., He, X.: SentiCap: generating image descriptions with sentiments. In: AAAI, pp. 3574–3580 (2016)
34. Mathews, A.P., Xie, L., He, X.: SemStyle: learning to generate stylised image captions using unaligned text. In: CVPR, pp. 8591–8600 (2018)
35. Nezami, O.M., Dras, M., Wan, S., Paris, C.: Senti-attend: image captioning using sentiment and attention. arXiv:1811.09789 (2018)

36. Mohamad Nezami, O., Dras, M., Wan, S., Paris, C., Hamey, L.: Towards generating stylized image captions via adversarial training. In: Nayak, A.C., Sharma, A. (eds.) PRICAI 2019. LNCS (LNAI), vol. 11670, pp. 270–284. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-29908-8_22
37. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: ACL, pp. 311–318 (2002)
38. Patterson, G., Hays, J.: COCO attributes: attributes for people, animals, and objects. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 85–100. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_6
39. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: CVPR, pp. 7008–7024 (2017)
40. Shin, A., Ushiku, Y., Harada, T.: Image captioning with sentiment terms via weakly-supervised sentiment dataset. In: BMVC (2016)
41. Vijayakumar, A.K., et al.: Diverse beam search: decoding diverse solutions from neural sequence models. arXiv:1610.02424 (2016)
42. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. In: CVPR, pp. 3156–3164 (2015)
43. Wang, L., Schwing, A., Lazebnik, S.: Diverse and accurate image description using a variational auto-encoder with an additive Gaussian encoding space. In: NIPS, pp. 5756–5766 (2017)
44. Xu, K., et al.: Show, attend and tell: neural image caption generation with visual attention. In: ICML, pp. 2048–2057 (2015)
45. Yao, T., Pan, Y., Li, Y., Mei, T.: Hierarchy parsing for image captioning. In: ICCV, pp. 2621–2629 (2019)
46. You, Q., Jin, H., Luo, J.: Image captioning at will: a versatile scheme for effectively injecting sentiments into image descriptions. arXiv:1801.10121 (2018)
47. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. TACL **2**, 67–78 (2014)
48. Zhao, W., Wu, X., Zhang, X.: MemCap: memorizing style knowledge for image captioning. In: AAAI, pp. 12984–12992 (2020)