# IMAGE CAPTIONING WITH INHERENT SENTIMENT

*Tong Li, Yunhui Hu, Xinxiao Wu**

Beijing Laboratory of Intelligent Information Technology
School of Computer Science, Beijing Institute of Technology, Beijing, China
{litong11, 1120183513, wuxinxiao}@bit.edu.cn

## ABSTRACT

We propose a new task called ***sentimental image captioning*** which aims to generate captions with the inherent sentiment reflected by the image. Compared with the stylized image captioning task that requires a predefined style independent of the image, our new task can automatically analyze the inherent sentiment tendency within the image. With this in mind, we propose an Inherent Sentiment Image Captioning (***InSenti-Cap***) method that first extracts the content and sentiment information from the image, and then fuses these information into the sentimental sentence generation via an attention mechanism. To effectively train the proposed model using the pairs of image and factual caption in existing captioning dataset and the extra sentiment corpus, we propose a two-stage training strategy that involves a sentimental regularization and a sentimental reward to enable the model to generate fluent and relevant sentences with inherent sentimental styles. Experiments demonstrate the effectiveness of our method.

***Index Terms***— Sentimental Image Captioning, Image Captioning, Image Sentiment Analysis

## 1. INTRODUCTION

The purpose of the stylized image captioning task is to generate image captions with a fixed or manually specified style and it has been explored in some works [1, 2, 3]. However, this task assumes that the linguistic style is predefined, which may not hold in real applications. Moreover, the given style may not consist with the underlying emotion of the image. For example, the emotion expressed in Figure 1(a) is happiness, so it is inappropriate to generate a specified negative caption. Hence, exploring the inherent sentiments within images is non-trivial and critical for generating more reasonable and sentimental image captions.

In this paper, we propose a new task, named sentimental image captioning, to generate an image caption that embodies the underlying sentiment expressed by the image. This new task relaxes the assumption of style independence in existing stylized image captioning methods, and has wide applications in real scenarios, such as helping people with visual impairments and infants in early education to better understand images from more perspectives and assisting social platforms to automatically generate appropriate captions for the images uploaded by users. However, the sentimental image captioning is very challenging since it not only needs to understand the image content, but also needs to analyze the intrinsic image sentiment and incorporate the sentimental elements into captioning. Moreover, there are no pairs of image and sentimental caption and the cost of collecting them is very expensive.


(a) A positive example. (b) A negative example.

**Fig. 1**. Examples that reflect positive and negative sentiments. (a) and (b) show a factual caption and a caption with the image sentiment, respectively.

To address the challenging issues, we propose an Inherent Sentiment Image Captioning (InSenti-Cap) method. It first extracts the content and sentiment information from the image, then fuses these information by an attention mechanism, and finally uses the fused information for sentimental caption generation. To be specific, we design three detectors, namely feature detector, sentiment detector and concept detector, to extract the visual features, sentiment category and concept words from the image, respectively. And we construct a prior knowledge base from the sentimental corpus to infer sentiment words related to the image according to the

*Corresponding Author.

This work was supported in part by the Natural Science Foundation of China (NSFC) under Grants No. 62072041.

978-1-6654-3864-3/21/$31.00 

extracted concept words. In the process of generating captions, the captioner introduces an attention module to decide whether to focus on the content or the sentiment of the image.

To effectively learn the sentimental image captioning model, we propose a two-stage training strategy. In the first stage, we train the captioner using the pairs of image and factual caption and the independent sentimental corpus. Inspired by self-supervised learning, we use a new sentimental regularization term to let the captioner learn how to add sentimental elements to the generated sentence, which is beneficial for faster and better training in the next stage. In the second stage, we integrate reinforcement learning to fine-tune the captioner. In addition to the commonly used CIDEr reward, we also propose a sentimental reward that encourages the captioner to pay more attention to the sentiment part of the sentence.

The main contributions of this paper are: (1) To the best of our knowledge, we are the first to propose the sentimental image captioning task that aims to generate image captions with the inherent sentiment reflected by the image. (2) We propose an InSenti-Cap method for the sentimental image captioning and a two-stage training strategy is proposed to incorporate the knowledge learned from the sentimental corpus into the caption generation process. (3) Experiments on the MSCOCO dataset validate the superior performance of our method.

## 2. RELATED WORK

### 2.1. Stylized Image Captioning

Recently, stylized image captioning has attracted increasing attention and several methods have been proposed. Mathews et al. [1] propose a switching RNN with word-level regularization, which can generate positive or negative captions. Guo et al.[2] utilize an adversarial learning network to handle multiple styles simultaneously. Zhao et al. [3] design a style memory module for memorizing the style knowledge learned from corpus. However, all these works generate a caption with a style unrelated to the image, while our method generates a caption with the inherent image sentiment.

### 2.2. Image Sentiment Analysis

As CNNs have achieved remarkable success in many computer vision tasks, it has also been used for image sentiment analysis. You et al. [4] employ CNNs to extract image features and then add several fully connected layers to recognize image sentiment. Yang et al. [5] not only use the global information of the image, but also consider the local information of the image. In this paper, we design an image sentiment detector based on the existing method [5], and we effectively improve its performance through threshold filtering on the sentiment scores.

## 3. OUR METHOD

### 3.1. Overview

The overall process of our method is as follows:

Firstly, we utilise a set of detectors to extract the content and sentiment information of images and sentimental corpus respectively. For the sentimental corpus, we employ the NLTK tool [6] to mark the part of speech of the sentences, and select the nouns and verbs as concept words. Moreover, object-sentiment word pairs are extracted as prior knowledge, which is used to acquire sentiment words based on the selected concept words. For images, we extract visual features $\mathcal{V}^c = \left\{\boldsymbol{v}_1^c, \boldsymbol{v}_2^c, \cdots, \boldsymbol{v}_{N_v}^c\right\}$ where $N_v$ is the number of visual features, concept words, sentiment catagory, and then obtain the global feature $\boldsymbol{v}^g = \frac{1}{N_v}\sum_{i=1}^{N_v} \boldsymbol{v}_i^c$ through the mean operation. In addition, sentiment words are acquired through the prior knowledge learned from sentimental corpus.

Subsequently, following [7], we combine two LSTMs and an attention module as our captioner, which generates captions based on the information extracted from images and reconstructs the sentimental sentences based on the information extracted from corpus.

Finally, we train our captioner in two stages. In the first stage, we pre-train the captioner using cross-entropy loss with the sentiment regularization term provided by the sentimental sentence reconstruction task. In the second stage, we fine-tune the model by adding reinforcement learning with a new sentimental reward calculated by a sentence sentiment classifier.

### 3.2. Content and Sentiment Extraction

The content and sentiment information of images and sentimental corpus will be used in the image caption generation task and the sentimental sentence reconstruction task. For the feature detector and the concept detector, we choose the off-the-shelf models from [8] and [9].

We choose the Detection Branch in [5] as the sentiment detector, because it achieves better result in the experiment. We train this model using the image sentiment analysis datasets. Then the detector is used to detect the sentiment $s$ of the image in the image captioning field. Since the image captioning dataset is mainly collected from daily scenes, it is mostly neutral, while the sentiments of the images in the image sentiment analysis field are more distinct. Therefore, we propose a threshold filtering on sentiment scores method to reduce the domains gap, that is, only when the sentiment score exceeds the threshold $\lambda_{ss}$, we consider the image to have this sentiment, otherwise it is considered neutral.

For the prior knowledge, we first obtain the correspondence between nouns and adjectives in the sentences from sentimental corpus. Then TF-IDF method is used to filter out adjectives that have nothing to do with sentiment. That is, when an adjective appears more frequently in sentences with

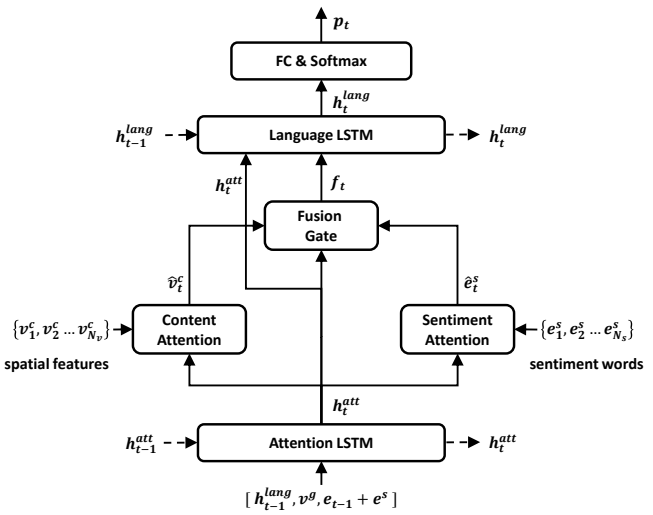



**Fig. 2**. Captioner. The captioner consists of an attention LSTM (bottom), an attention module (middle) and a language LSTM (top). It is used to complete image caption generation and sentimental sentence reconstruction tasks based on the information extracted by the detectors.

a certain sentiment and less frequently in all sentences, we consider it to be a sentiment word, otherwise it is not. Now, we get the prior knowledge base and we can utilize it to extract the sentiment words corresponding to the objects in the concept words.

### 3.3. Sentimental Image Captioner

The captioner is used for image caption generation and sentimental sentence reconstruction. The image caption generation task takes images as input and generates sentimental captions. The sentimental sentence reconstruction task reconstructs the sentences in the sentimental corpus through self-supervised learning, and its purpose is to provide a regularization term for the image caption generation task, which is helpful for the captioner to learn sentimental knowledge from the corpus.

As shown in Figure 2, the captioner includes an attention LSTM, an attention module and a language LSTM. The function of the attention LSTM is to guide the attention module on which piece of information extracted by the detectors should be currently focused on. The attention module delivers the fusion features of content and sentiment to the language LSTM, and the language LSTM generates a sentence word by word.

#### *3.3.1. Attention LSTM*

At each time step, its input includes the previous output $\boldsymbol{h}_{t-1}^{lang}$ of the language LSTM, concatenated with the global feature $\boldsymbol{v}^g$ and the sum of the embedding vector $\boldsymbol{e}_{t-1}$ of the word generated in previous step and the embedding vector $\boldsymbol{e}^s$ of the image sentiment $s$, expressed as $[\boldsymbol{h}_{t-1}^{lang}; \boldsymbol{v}^g; \boldsymbol{e}_{t-1} + \boldsymbol{e}^s]$. Note that we first convert the extracted visual features into the same dimension as the word embedding vector through linear transformation.

For sentimental sentence reconstruction, the image sentiment $s$ is replaced by the sentiment label of the sentence, and because there is no image data, we map the concept words features to the visual global feature space for simulation. First, we represent the concept words extracted from the image and selected from the sentimental sentence as the corresponding word embedding vectors $\mathcal{E}^{ic} = \left\{\boldsymbol{e}_1^{ic}, \boldsymbol{e}_2^{ic}, \cdots, \boldsymbol{e}_{N_c}^{ic}\right\}$ and $\mathcal{E}^{sc} = \left\{\boldsymbol{e}_1^{sc}, \boldsymbol{e}_2^{sc}, \cdots, \boldsymbol{e}_{N_c}^{sc}\right\}$ respectively, where $N_c$ is the number of concept words. Then we obtain their global feature $\boldsymbol{e}^{ig} = \frac{1}{N_c}\sum_{j=1}^{N_c} \boldsymbol{e}_j^{ic}$ and $\boldsymbol{e}^{sg} = \frac{1}{N_c}\sum_{j=1}^{N_c} \boldsymbol{e}_j^{sc}$ through the mean operation. Next, map them to the visual global feature space: $\boldsymbol{v}_{ic}^g = \boldsymbol{W}_{cg}\boldsymbol{e}^{ig}, \boldsymbol{v}_{sc}^g = \boldsymbol{W}_{cg}\boldsymbol{e}^{sg}$. Finally, we learn the parameter $\boldsymbol{W}_{cg}$ by reducing the mean square error between $\boldsymbol{v}_{ic}^g$ and $\boldsymbol{v}^g$, and use $\boldsymbol{v}_{sc}^g$ to simulate $\boldsymbol{v}^g$ during sentimental sentence reconstruction.

#### *3.3.2. Attention Module*

First, we convert sentiment words into word embedding vectors $\mathcal{E}^s = \left\{\boldsymbol{e}_1^s, \boldsymbol{e}_2^s, \cdots, \boldsymbol{e}_{N_s}^s\right\}$, where $N_s$ is the number of sentiment words. Then, we utilise the content attention module to fuse visual features $\mathcal{V}^c$ and the sentiment attention module to fuse sentiment words features $\mathcal{E}^s$. Finally, the gating mechanism is used to fuse content and sentiment features.

For the content attention module, at time step $t$, given the hidden state $\boldsymbol{h}_t^{att}$ of the attention LSTM, we generate attention weight $\alpha_{i,t}^v$ for each visual feature $\boldsymbol{v}_i^c$, as follows:

$$\begin{aligned} a_{i,t}^v &= \boldsymbol{W}_a^{vT} \tanh(\boldsymbol{W}_{va}\boldsymbol{v}_i^c + \boldsymbol{W}_{ha}^v \boldsymbol{h}_t^{att}), \\ \boldsymbol{\alpha}_t^v &= \text{softmax}(\boldsymbol{a}_t^v), \end{aligned} \tag{1}$$

where $\boldsymbol{W}_{va} \in \mathbb{R}^{H\times E}$, $\boldsymbol{W}_{ha}^v \in \mathbb{R}^{H\times A}$, and $\boldsymbol{W}_a^v \in \mathbb{R}^H$ are learnable parameters. Then the features are fused by the following formula:

$$\hat{\boldsymbol{v}}_t^c = \sum_{i=1}^{N_v} \alpha_{i,t}^v \boldsymbol{v}_i^c. \tag{2}$$

For the sentiment attention module, given $\boldsymbol{h}_t^{att}$ and sentiment vector $\boldsymbol{e}^s$, attention weight $\alpha_{j,t}^s$ of each sentiment word $\boldsymbol{e}_j^s$ is calculated by:

$$\begin{aligned} a_{j,t}^s &= \boldsymbol{W}_a^{sT} \tanh(\boldsymbol{W}_{ea}\boldsymbol{e}_j^s + \boldsymbol{W}_{sa}\boldsymbol{e}^s + \boldsymbol{W}_{ha}^s \boldsymbol{h}_t^{att}), \\ \boldsymbol{\alpha}_t^s &= \text{softmax}(\boldsymbol{a}_t^s), \end{aligned} \tag{3}$$

where $\boldsymbol{W}_{ea}, \boldsymbol{W}_{sa} \in \mathbb{R}^{H\times E}$, $\boldsymbol{W}_{ha}^s \in \mathbb{R}^{H\times A}$ and $\boldsymbol{W}_a^s \in \mathbb{R}^H$ are learnable parameters. And the fusion process of sentiment features $\hat{\boldsymbol{e}}_t^s$ is similar to Formula 2.

Finally, we introduce fusion gate to fuse content and sentiment features:

$$\begin{aligned} b_t &= \boldsymbol{W}_f{}^T \tanh(\boldsymbol{W}_{cb}\hat{\boldsymbol{v}}_t^c + \boldsymbol{W}_{sb}\hat{\boldsymbol{e}}_t^s + \boldsymbol{W}_{hb}\boldsymbol{h}_t^{att}), \\ \beta_t &= \text{sigmoid}(b_t), \boldsymbol{f}_t = \beta_t \hat{\boldsymbol{v}}_t^c + (1-\beta_t)\hat{\boldsymbol{e}}_t^s, \end{aligned} \tag{4}$$

where $\boldsymbol{W}_{cb}, \boldsymbol{W}_{sb} \in \mathbb{R}^{H\times E}$, $\boldsymbol{W}_{hb} \in \mathbb{R}^{H\times A}$ and $\boldsymbol{W}_f \in \mathbb{R}^H$ are learnable parameters.

For sentimental sentence reconstruction task, we ignore the content attention module and directly use $\hat{\boldsymbol{e}}_t^s$ as the final output $\boldsymbol{f}_t$.

#### 3.3.3. Language LSTM

The language LSTM takes $[\boldsymbol{f}_t; \boldsymbol{h}_t^{att}]$ as input to generate the current word. And the conditional distribution over possible output words is calculated through a linear transformation and a softmax operation.

Notice that, for simplicity, we omit all the bias for linear transformations described above.

### 3.4. Sentence Sentiment Classification

We design a sentence sentiment classifier that provides rewards for reinforcement learning and is also used as a final sentiment evaluation metric.

A LSTM is used to encode each word in the sentence: $\boldsymbol{h}_t = \text{LSTM}(e_t)$, where $\boldsymbol{e}_t$ represents the word embedding vector, and $\boldsymbol{h}_t$ represents the encoding result.

Because different parts of the sentence have different effects on the overall sentiment classification, we perform a squeeze-excitation operation [10] to strengthen important encoding features. The enhanced features are fused as the final sentence features. And the sentimental probability is output through a linear transformation and a softmax operation. Due to the different sentence lengths, we make a simple modification to the squeeze-excitation operation: first perform the excitation operation and then perform the squeeze operation. The entire operation process is as follows:

$$\begin{aligned}
\boldsymbol{m}_t &= \text{excitation}(\boldsymbol{h}_t) = \text{sigmoid}(\boldsymbol{W}_2(\text{Relu}(\boldsymbol{W}_1\boldsymbol{h}_t))),\\
\boldsymbol{\alpha} &= \text{squeeze}(\boldsymbol{M}) = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{m}_t,\\
\boldsymbol{f}_s &= \text{fuse}(\boldsymbol{\alpha}, \boldsymbol{H}) = \sum_{t=1}^{T}\alpha_t\boldsymbol{h}_t,\\
\boldsymbol{p}_s &= \text{softmax}(\boldsymbol{W}_3\boldsymbol{f}_s),
\end{aligned} \tag{5}$$

where $\boldsymbol{W}_1, \boldsymbol{W}_2 \in \mathbb{R}^{A\times A}$ and $\boldsymbol{W}_3 \in \mathbb{R}^{K\times A}$ are learnable parameters and K is the number of sentiment catagories.

The model is trained using the cross-entropy loss. When joining reinforcement learning training the captioner, we use the importance score $\boldsymbol{\alpha}$ as the reward, which helps the captioner pay attention to the sentimental part of the sentence.

### 3.5. Train Strategy

#### 3.5.1. Pre-training stage

At this stage, in addition to the standard cross-entropy loss $\mathcal{L}_{XE}$ and the mean square error $\mathcal{L}_{da}$ mentioned above, we also employ the sentimental sentence reconstruction loss $\mathcal{L}_{re}$ as a regularization term. When generating the factual caption, we ignore the sentiment attention module and use the sentiment of the ground-truth captions (classified by the sentence sentiment classifier) instead of the image sentiment $s$ as the input of the attention LSTM. The loss at this stage is expressed as:

$$\begin{aligned}
\mathcal{L}_{Pt} &= \mathcal{L}_{XE} + \mathcal{L}_{da} + \mathcal{L}_{re},\\
\mathcal{L}_{XE} &= -\frac{1}{T}\sum_{t=1}^{T}\log(p_t(y_t^f|y_{1:t-1}^f)),\\
\mathcal{L}_{da} &= \text{MSE}(\boldsymbol{v}_{ic}^g, \boldsymbol{v}^g),\\
\mathcal{L}_{re} &= -\frac{1}{T}\sum_{t=1}^{T}\log(p_t(y_t^s|y_{1:t-1}^s)),
\end{aligned} \tag{6}$$

where $p_t(y_t|y_{1:t-1})$ denotes the predicted probability of the ground-truth word $y_t$ given the previous word sequence $y_{1:t-1}$.

#### 3.5.2. Fine-tuning stage

At this stage, besides the above loss $\mathcal{L}_{Pt}$, we also add reinforcement learning. The gradient of the captioner parameters $\theta$ is approximated by:

$$\nabla_\theta\mathcal{L}_R(\theta) \approx -r(y^s, \hat{y})\nabla_\theta\log\theta(y^s), \tag{7}$$

where $y^s$ is a sampled caption and $\hat{y}$ denotes the sentence generated by greedy decoding. The components of the reward function $r$ are:

$$\begin{aligned}
r(y^s, \hat{y}) &= \lambda_1 r_{CIDEr} + \lambda_2 r_{cls},\\
r_{CIDEr} &= \text{CIDEr}(y^s) - \text{CIDEr}(\hat{y}),\\
r_{cls} &= \mathbb{I}_{(s_i=s_s)}\boldsymbol{\alpha}_{y^s},
\end{aligned} \tag{8}$$

where $\text{CIDEr}(y)$ is the CIDEr score of sentence $y$, $\boldsymbol{\alpha}_y$ represents the importance score calculated by the sentence sentiment classifier for sentence $y$, and if the image sentiment detection result $s_i$ is consistent with the sentence sentiment classification result $s_s$, $\mathbb{I}_{(s_i=s_s)}$ takes 1 otherwise it takes 0.

## 4. EXPERIMENTS

### 4.1. Datasets

**Image captioning dataset.** We choose MSCOCO [11] dataset. And we use the Karpathy splits [12] for the model validation and offline evaluation. In this split, 113287 and 5000 images with five factual captions are for training and validation, respectively. 5000 images are for test.

**Image sentiment dataset.** We select the Emotion-ROI [13], ArtPhoto [14], Twitter I [4] and Twitter II [15] datasets in the image sentiment analysis field for training the sentiment detector. For Twitter I, we use the "At Least Four

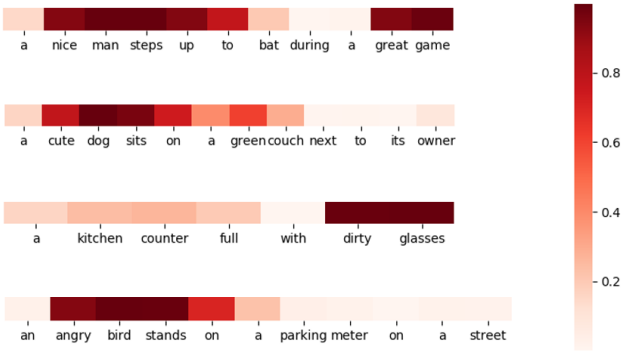



**Fig. 3**. Visualization examples of the importance scores $\boldsymbol{\alpha}$ of each part of the sentence in the process of sentiment classification. The redder is more important.

**Table 1**. Comparison with the stylized image captioning methods on MSCOCO dataset. B@n, M and C are the abbreviations of Bleu-n, METEOR and CIDEr respectively. * indicates that a single model can only generate captions of one sentiment. For ppl metric, the smaller value is better, and for other metrics, the larger value is better.

| Method | Sentiment | B@1 | B@3 | M | C | ppl(↓) | cls(%) |
|---|---|---|---|---|---|---|---|
| MemCap* | positive | 50.8 | 17.1 | 16.6 | 54.4 | 13.0 | 99.8 |
| | negative | 48.7 | 19.6 | 15.8 | 60.6 | 14.6 | 93.1 |
| MSCap | positive | 46.9 | 16.2 | 16.8 | 55.3 | 19.6 | 92.5 |
| | negative | 45.5 | 15.4 | 16.2 | 51.6 | 19.2 | 93.4 |
| MemCap | positive | 51.1 | 17.0 | 16.6 | 52.8 | 18.1 | 96.1 |
| | negative | 49.2 | 18.1 | 15.7 | 59.4 | 18.9 | 98.9 |
| InSenti-Cap | positive | 59.7 | 25.3 | 20.9 | 61.3 | 13.0 | 98.5 |
| | negative | 59.1 | 24.3 | 19.4 | 53.3 | 12.3 | 95.5 |
| | neutral | 73.5 | 41.2 | 24.7 | 97.5 | 8.4 | 98.9 |

Agree" result which indicates that at least 4 AMT workers gave the same sentiment label for a given image. And the EmotionROI and ArtPhoto datasets contain multiple sentiments. In this paper, we only focus on positive and negative sentiments, so we reclassify these images into positive, negative and neutral categories.

**Sentiment corpus.** We employ the SentiCap [1] dataset, which includes 4892 positive sentences and 3977 negative sentences. For the neutral category, we select the sentences from image captioning dataset that do not contain sentiment words (from the prior knowledge) to expand.

### 4.2. Implementation Details

We extract the 2048-dimensional visual features of images through the last convolutional layer of the trained ResNet-101 [8]. The dimensions of all LSTMs' hidden states and the size of the word embedding vector are set to 512. The numbers of concept words $N_c$ and sentimental words $N_s$ are set to 5 and 10 respectively. The two parameters $\lambda_1$ and $\lambda_2$ in Equation 8 are set to 1 and 0.6, respectively. In the pre-training stage, the learning rate is set to $4\times10^{-4}$, and in the fine-tuning stage, the learning rate is set to $4 \times 10^{-5}$. We employ the Adam optimizer [16] to train all models. Source code is available at `https://github.com/ezeli/InSentiCap_model`.

### 4.3. Results

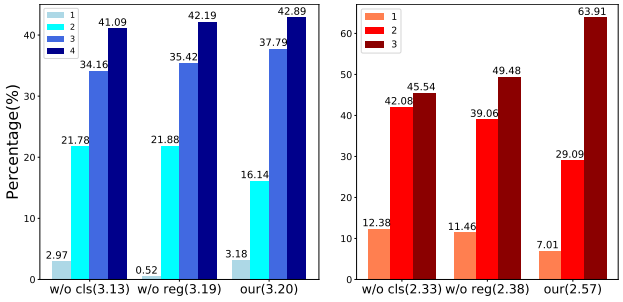
**Sentiment Detector Performance.** To evaluate the performance of the sentiment detector, we collect a new image sentiment dataset based on the MSCOCO dataset, in which the numbers of positive, negative and neutral sentiment images are 81, 35, and 137. The experimental result is that if the detector is directly used (that is, the hyperparameter $\lambda_{ss}$ is set to 0), the detector's accuracy is only 62.9%, but when the $\lambda_{ss}$ is set to 0.7, the accuracy is improved to 65.6%. This shows that threshold filtering on sentiment scores is effective in solving the domain gap between the images in the image captioning field and the image sentiment analysis field.

**Sentence Sentiment Classifier Performance.** We train the sentence sentiment classifier using sentiment corpus and achieve 99% accuracy. Figure 3 shows several visualization examples of the importance scores $\boldsymbol{\alpha}$ of each part of the sentence when sentiment classification. We can see that the classifier can accurately capture the sentimental part of the sentence, which is conducive to providing useful rewards in reinforcement learning for generating sentimental captions.

**Comparison with stylized image captioning methods.** In order to evaluate the quality of generated sentiment captions, we compare with the state-of-the-art methods in the stylized image captioning field, including MSCap [2] and MemCap [3], as shown in Table 1. The model with the symbol * in the table trains one model for each style, and other models train one model for multiple styles. Following [3], to verify the content relevance, we report the widely used automatic evaluation metrics, i.e., BLEU, METEOR and CIDEr. And we measure sentiment consistency through the perplexity score calculated by the language model (denoted as ppl) and the sentiment classification accuracy classified by the sentence sentiment classifier (denoted as cls). Our InSenti-Cap generates multiple sentimental captions using a single model and utilizes the sentiment detection result as the corresponding sentiment.

From Table 1, we can have the observations as follows: (1) Compared with MSCap and MemCap, our model has a very significant improvement in most metrics, which illustrates that our model can generate captions that are more in line with the content of the image and the corresponding sentiment. (2) Compared with MemCap*, our method also outperforms in most metrics, validating the superiority of our method on capturing multiple sentiment knowledge for captioning.

**Ablation Studies.** The automatic evaluation metrics can-




(a) Content Relevance. (b) Sentiment Consistency.

**Fig. 4**. Ablation studies on the MSCOCO dataset through human evaluation. The content relevance metric is rated from 1 (unrelated) to 4 (very related), and the sentiment consistency metric is rated from 1 (bad) to 3 (perfect). The vertical axis represents the percentage of each score, and the average score is in parentheses.

not reflect the quality of the generated captions very well, so we perform human evaluation to conduct ablation studies, and its results are reported in Figure 4. Specifically, during the entire training process, we remove the sentimental regularization term $\mathcal{L}_{re}$, denoted as "w/o reg". In the fine-tuning stage, we remove the sentimental reward function $r_{cls}$, denoted as "w/o cls". For the human evaluation, we first randomly select 150 images and generate captions. Then, seven volunteers are invited to conduct quality assessment. The content relevance is rated from 1 (unrelated) to 4 (very related), and the sentiment consistency is rated from 1 (bad) to 3 (perfect). From the Figure 4, we can see that compared with "w/o reg" and "w/o cls", the content relevance metric and the sentiment consistency metric of our method have been improved, which validates that both the sentimental regularization and the sentimental reward function promote the model to focus on the sentimental part of the generated captions.

## 5. CONCLUSION

We have proposed a novel sentimental image captioning task that can generate captions that are more in line with image sentiment. We also have represented an InSenti-Cap method for this task and designed a two-stage training strategy to learn our model using the pairs of image and factual caption and the extra sentiment corpus. Our model is capable of understanding the content and sentiment of the image, and generate a caption with the image sentiment simultaneously. Moreover, experiments demonstrate the superiority of our method.

## 6. REFERENCES

[1] Alexander Mathews, Lexing Xie, and Xuming He, "Senticap: Generating image descriptions with sentiments," in *AAAI*, 2016, pp. 3574–3580.

[2] Longteng Guo, Jing Liu, Peng Yao, Jiangwei Li, and Hanqing Lu, "Mscap: Multi-style image captioning with unpaired stylized text," in *CVPR*, 2019, pp. 4204–4213.

[3] Wentian Zhao, Xinxiao Wu, and Xiaoxun Zhang, "Memcap: Memorizing style knowledge for image captioning," in *AAAI*, 2020, pp. 12984–12992.

[4] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *AAAI*, 2015, pp. 381–388.

[5] Jufeng Yang, Dongyu She, Yu-Kun Lai, Paul L Rosin, and Ming-Hsuan Yang, "Weakly supervised coupled networks for visual sentiment analysis," in *CVPR*, 2018, pp. 7584–7592.

[6] Steven Bird, Ewan Klein, and Edward Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*, O'Reilly Media, 2009.

[7] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *CVPR*, 2018, pp. 6077–6086.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[9] Weixuan Wang, Zhihong Chen, and Haifeng Hu, "Hierarchical attention network for image captioning," in *AAAI*, 2019, pp. 8957–8964.

[10] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018, pp. 7132–7141.

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014, pp. 740–755.

[12] Andrej Karpathy and Li Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *CVPR*, 2015, pp. 3128–3137.

[13] Kuan-Chuan Peng, Amir Sadovnik, Andrew Gallagher, and Tsuhan Chen, "Where do emotions come from? predicting the emotion stimuli map," in *ICIP*, 2016, pp. 614–618.

[14] Jana Machajdik and Allan Hanbury, "Affective image classification using features inspired by psychology and art theory," in *ACM MM*, 2010, pp. 83–92.

[15] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *ACM MM*, 2013, pp. 223–232.

[16] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.